

DAS SYSTEM
ANCA ZUR CLUSTERANALYSE -
EIN BEITRAG ZUR INTEGRATION MULTIVARIATER
VERFAHREN IN EINE METHODENBANK

VON
DIPL.-KFM. KARL SCHALLER
AUS
NAILA

1979



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41546213_0001

DAS SYSTEM ANCA ZUR CLUSTERANALYSE - EIN BEITRAG ZUR INTEGRATION MULTIVARIATER VERFAHREN IN EINE METHODENBANK

INAUGURAL - DISSERTATION

ZUR ERLANGUNG DES AKADEMISCHEN GRADES
EINES DOKTORS DER WIRTSCHAFTS- UND SOZIALWISSENSCHAFTEN
(DR. RER. POL.)

AN DER WIRTSCHAFTS- UND SOZIALWISSENSCHAFTLICHEN FAKULTÄT
DER FRIEDRICH - ALEXANDER - UNIVERSITÄT ERLANGEN - NÜRNBERG

VORGELEGT

VON

DIPL.-KFM. KARL SCHALLER

AUS

NAILA

DAS SYSTEM

ANCA ZUR CLUSTERANALYSE

EIN BEITRAG ZUR INTEGRATION MULTIVARIATER
VERFAHREN IN EINE METHODOLOGIE



Referent: Prof. Dr. P. Mertens

Korreferent: Prof. Dr. M. Meyer

Nürnberg, 18. 12. 1979

1. Zielsetzung der Arbeit	1
1.1. Überblick	1
1.2. Einordnung in die Entwicklungstrends für Methodenbanken	3
1.3. Methodenspektrum von Methodenbanken	6
1.4. Zielsetzung beim Einbau der Clusteranalyse in eine Methodenbank	8
2. Ausgangsbasis	13
2.1. Programmsysteme zur Clusteranalyse	13
2.1.1. Isolierte Programme	13
2.1.2. Spezielle Programmsysteme zur Clusteranalyse bei Verlagen und Instituten	14
2.1.3. Generelle Programmsysteme zur Clusteranalyse	15
2.1.4. Clusteranalyse in Standard-Programmsystemen	17
2.2. Die interaktive Methodenbank MADAS	18
2.2.1. Zielgruppe	18
2.2.2. Verknüpfung mit der Datenbank	20
2.2.3. Methodenspektrum in MADAS	21
2.2.4. Verarbeitungsgrenzen	23
2.3. Methoden der Clusteranalyse in MADAS	24
2.3.1. Vorteile der Integration für MADAS	24
2.3.2. Vorteile der Integration für die Clusteranalyse	26
2.3.3. ANCA als selbständige Teilmenge von MADAS	29
3. Die Methodenbank ANCA aus der Sicht des Anwenders	32
3.1. Systembasis von ANCA	32
3.1.1. Methodensammlung	32
3.1.1.1. Reichtum des Inhalts	32
3.1.1.2. Erweiterbarkeit der Methodensammlung	40
3.1.1.3. Richtigkeit der Methodensammlung	42
3.1.2. Methodendokumentation	43
3.1.3. Interaktive Verarbeitung	44
3.1.4. Training	49
3.2. Systemablauf bei einer Clusteranalyse	50
3.2.1. Daten- und Merkmalsauswahl	50
3.2.2. Methodenangebot	52
3.2.3. Methodenauswahl	55
3.2.3.1. Erscheinungsformen der Methodenauswahl	55
3.2.3.2. Auswahlhilfen für Verfahren der Datenvorbehandlung	60
3.2.3.3. Auswahlhilfen für Abstandsmaße	63
3.2.3.4. Auswahlhilfen für Clusteralgorithmen	65
3.2.3.5. Auswahlhilfen für partitionierende Verfahren	68
3.2.4. Methodenverknüpfung	69
3.2.5. Parameterversorgung	70
3.2.6. Methodendurchlauf	72
3.2.7. Interpretation der Ergebnisse	73
3.2.7.1. Allgemeines	73
3.2.7.2. Analyse der Hierarchie	74
3.2.7.3. Ergebnisdarstellung für die Cluster	81
3.2.7.4. Bewertung der Cluster	88
3.2.8. Informationshilfen	94

3.3. Systemablauf bei mehreren (vergleichenden) Analysen	96
3.3.1. Anwendungsmöglichkeiten	96
3.3.2. Varianten des Ergebnisvergleichs in ANCA	98
3.3.3. Dialogablauf bei vergleichender Analyse	101
3.4. Systemablauf bei Methodenverbund in MADAS	105
4. Datenverarbeitungstechnische Realisation des Systems ANCA	111
4.1. Grundkonzept	111
4.1.1. Bedarfsangaben	111
4.1.2. Aufruf des Systems	112
4.1.3. Dialog am Graphikbildschirm	112
4.2. Programmaufbau von ANCA	114
4.2.1. Sprachkonzept	114
4.2.2. Modulkonzept	115
4.2.3. Steuerkonzept	117
4.2.4. Dateikonzept	118
4.2.5. Schnittstellenkonzept	120
4.3. Dialogablauf	121
4.3.1. Modulorientierter Durchlauf	121
4.3.2. Kommandoorientierte Steuerung	121
5. Kritische Analyse der Entwicklung und Anregungen für zukünftige Aktivitäten	125
5.1. Diskussion des Gesamtkonzepts	125
5.1.1. Zielgruppendifkussion	125
5.1.2. Diskussion der interaktiven Verarbeitung	127
5.1.3. Möglichkeiten Graphischer Datenverarbeitung	130
5.2. Diskussion des Methodenbereichs	131
5.2.1. Methodensammlung	131
5.2.2. Methodenwahl	134
5.2.2.1. Forderung nach einer verbesserten Methodenwahl	134
5.2.2.2. Vergleiche von Verfahren zur Clusteranalyse in der Literatur	136
5.2.2.2.1. Vergleiche von hierarchischen Verfahren	137
5.2.2.2.2. Sonstige Vergleiche	139
5.2.2.3. Methodenwahl als Mustererkennungsansatz	142
5.2.2.3.1. Kriterienkatalog	142
5.2.2.3.2. Gestaltung des Mustererkennungsprozesses	147
5.2.3. Ergebnisbehandlung	150
5.3. Diskussion des DV-Konzepts	151
5.3.1. Offenheit des Systems	151
5.3.2. Dialogablauf	152
5.3.3. Einbau von Benutzermodellen	154
5.4. Diskussion des Umfeldes von ANCA	155
5.4.1. Verbund von Systemen zur Clusteranalyse	155
5.4.2. Ergänzung um Hilfsroutinen	156
6. Ausblick	158
Literaturverzeichnis	

1. Zielsetzung der Arbeit

1.1. Überblick

Die Forderung nach Methodenbanken zur Auswertung der immer stärker wachsenden Datenmengen war lange Zeit rein theoretischer Natur¹⁾. In der Zwischenzeit wurden in den verschiedensten Anwendungsbereichen - neben den hier betrachteten Wirtschaftswissenschaften²⁾ vor allem in der Medizin³⁾ - erste Pilotprojekte verwirklicht und sind auf positive Resonanz gestoßen. In bezug auf Zielgruppenkonzept, Entwurfskriterien und datenverarbeitungstechnische Realisierung sind diese Ansätze sehr uneinheitlich; Kritiker meinten deshalb noch vor kurzer Zeit, Methodenbanken seien "mehr Vorstellung als Realität"⁴⁾ bzw. schrieben, die Diskussion über sie "bewegt sich ... weitgehend auf spekulativer Ebene"⁵⁾.

1) Vgl. z. B. Breitung, A., Marketing-Informationssysteme, Meisenheim 1975.

2) Vgl. z. B. zu METHAPLAN: Esprester, A. C., Die Entwicklung einer Methodenbank und einer Methodensprache, Angewandte Informatik 20 (1978), S. 203 ff.; Rickert, R., Methodenbanken - Softwaretechnologie für individuelle Planungs-, Informations- und Auswertungssysteme, Siemens-Zeitschrift 51 (1977), S. 242 ff.; vgl. zur Methodenbank im Marketing-Informationssystem der Henkel KGaA: Krautter, J. und Scholz, P., Entwicklung computergestützter Marketing-Informationssysteme, Die Betriebswirtschaft 38 (1978), S. 231 ff.; Cabus, E., Ansätze für ein Marketing-Informationssystem, in: Haedrich, G. (Hrsg.), Operationale Entscheidungshilfen für die Marketingplanung, Berlin-New York 1977, S. 269 ff.; vgl. zur ökonomischen Methodenbank MEBA: Schips, B., Ein Beitrag zum Thema "Methodenbanken", Angewandte Informatik 19 (1977), S. 465 ff.

3) Vgl. dazu die Veröffentlichungen im Statistical Software Newsletter, der vom Institut für Medizinische Datenverarbeitung der Gesellschaft für Strahlen- und Umweltforschung herausgegeben wird.

4) Samelson, K., Entwicklungslinien der Informatik, in: Schindler, S. und Giloi, W. K. (Hrsg.), Proceedings der Jahrestagung 1978 der Gesellschaft für Informatik, Berlin-Heidelberg-New York 1978, S. 142.

5) Heinzlbecker, K., Partielle Marketing-Informationssysteme, Moduln für computergestützte Entscheidungen in Marketing, Verkauf, Vertrieb, Zürich-Frankfurt 1977, S. 235.

In dieser Situation hat die Forschung zwei Aufgaben zu erfüllen:

- Sie muß nachweisen, daß Methodenbankkonzepte erfolgreich sein können.

Bekannte und neue Pilotprojekte müssen anhand von praktischen Problemstellungen ihre Leistungsfähigkeit und ihre Akzeptanzchancen zeigen, wobei die Ansätze bewertet und verglichen werden müssen.

Ziel dieser Forschung ist es, Entwurfskriterien für Methodenbanken herauszubilden und zu überprüfen.

- Sie muß die Möglichkeit schaffen, Methodenbanken für spezifische Aufgabenstellungen wirtschaftlich zu erstellen.

Es sind software-technologische Instrumente zu schaffen, die bei der Konzeption von Methodenbanken für veränderte oder erweiterte Aufgabenstellungen Wirtschaftlichkeit und Verwirklichung der o. a. Entwurfskriterien garantieren.

Insgesamt scheinen die Akzeptanzchancen für Methodenbanken noch nicht ausreichend erforscht. Unsere Forschungsgruppe konzentriert sich deshalb stärker auf die Erstellung und Anwendung von Pilotsystemen¹⁾;

¹⁾ Vgl. z. B.: Mertens, P. u.a., Verknüpfung von Daten- und Methodenbanken, dargestellt am Beispiel der Analyse von Marktforschungsdaten, in: Plötzeneder, H. D. (Hrsg.), Computergestützte Unternehmensplanung - Computer Assisted Corporate Planning, Stuttgart 1977, S. 291 ff.; Mertens, P. und Bodendorf, F., Interaktiv nutzbare Methodenbanken - Entwicklungskriterien und Stand der Verwirklichung, Angewandte Informatik (in Vorbereitung); Bodendorf, F., Renninger, W. und Schaller, K., Zwei neue Methodenbanksysteme für statistische Anwendungen, Arbeitsbericht des Instituts für Mathematische Maschinen und Datenverarbeitung 11 (1978) 22; Eckardt, Th., Eine Prognosemethodenbank für Kleinrechner - Konzept, Stand der Entwicklung und erste Erfahrungen, Vortrag anlässlich der 1. Sitzung der Arbeitsgruppe "Prognoseverfahren" der Deutschen Gesellschaft für Operations Research, Berlin 1978.

dabei stehen interaktive Systeme im Vordergrund, weil sie besser am Arbeitsplatz verfügbar und so als "Entscheidungs-labor" im Vorteil sind.

In der vorliegenden Arbeit wird die Integration von Verfahren der Clusteranalyse in eine Methodenbank untersucht. Neben der Implementierung des Anwendungs-orientierten Systems zur Clusteranalyse (ANCA) werden Hinweise für den Einbau "komplizierter" Verfahren in Methodenbanken geliefert.

1.2. Einordnung in die Entwicklungstrends für Methodenbanken

Einfachste Form von Methodenbanken sind Methodensammlungen, die dem Benutzer zur Verfügung gestellt werden, entweder in Form von Unterprogrammen¹⁾, für die noch geeignete Rahmenprogramme zu erstellen sind, oder als Stand-alone-Programme²⁾, die über Steuerkarten gestartet und mit Parametern versorgt werden müssen. Beide Varianten sorgen für die Bereitstellung der Verfahren, verhindern aber keinen Methodenmißbrauch und werden deshalb gemeinhin von Methodenspezialisten abgelehnt.

¹⁾ Vgl. z. B. IBM (Hrsg.), SSP III - Programmer's Manual, IBM-Form 6H 20-0205-4; Smith, B. T. e.a., Matrix Eigensystem Routines - EISPACK Guide, Lecture Notes in Computer Science 6, Berlin-Heidelberg-New York 1974.

²⁾ Vgl. z. B. Nie, N. H., Bent, D. H., Hull, O. H., Statistical Package for the Social Sciences, New York 1975; Dixon, W. J. (Hrsg.), BMD - Biomedical Computer Programs, Berkeley-Los Angeles 1970; derselbe, Biomedical Computer Programs - P-Series, Berkeley-Los Angeles 1977; ICPR (Hrsg.), OSIRIS III Manual, Ann Arbor 1974.

Neben Verbesserungen im Aufruf der Verfahren¹⁾ werden deshalb in neueren Entwicklungen Anstrengungen unternommen, die Auswahl der geeigneten Methode zu unterstützen²⁾.

Einigkeit herrscht in allen Ansätzen darüber, daß die Methoden in einfacher und verständlicher Form zur Verfügung gestellt werden müssen³⁾. Das erforderliche Ausmaß an Unterstützung des Anwenders wird dagegen unterschiedlich beurteilt. Hauptursache dafür ist die jeweils angewandte Definition der Zielgruppe, insbesondere die Frage, wie stark Entwickler und Anwender der Methodenbank getrennt sind:

1. Für den Fall, daß beide Funktionen in einer Hand liegen, z. B. bei einer Serviceabteilung, die speziell den "Methoden-Support" durchführt, oder bei Anwendern, die - beispielsweise in der Ökonometrie - überwiegend mit der Entwicklung von Modellen beschäftigt sind, ist vielfach ein geringer Komfort ausreichend. In der Regel genügt es, eine dialogorientierte "Arbeitsumgebung" zu schaffen. Grundlage können Sprachen wie APL⁴⁾ oder spezielle Systeme wie MEBA⁵⁾ sein; sie ermöglichen

¹⁾ Vgl. z. B. zur Entwicklung von METHAPLAN für die Methodenbank MEB Esprester, A. C., Die Entwicklung ..., a.a.O.

²⁾ Vgl. Mertens, P. u.a., Verknüpfung von ..., a.a.O.; Erbe, R. und Walch, G., An Interactive Guidance System for Method Libraries, IBM Publikation 75.004.001, Heidelberg 1975.

³⁾ Vgl. z. B. die empirische Untersuchung von Köhler und Uebele: Köhler, R. und Uebele, H., Planung und Entscheidung im Absatzbereich industrieller Großunternehmen, Aachen 1977, S. 173 ff.

⁴⁾ Vgl. Pakin, S., APL-Handbuch, Stuttgart 1975.

⁵⁾ Vgl. Schips, B., Ein Beitrag ..., a.a.O. und die dort angegebene Literatur; vgl. auch die Diskussion über modellgenerierende Software bei: Dickhoven, S. (Hrsg.), Modellierungssoftware, St. Augustin 1976.

sprach- und systemkundigen Anwendern, in kurzer Zeit ad-hoc-Modelle zu entwickeln, durchzurechnen und bei Bedarf abzuspeichern. Für komplizierte Verfahren können Schnittstellen zu fertigen Programmpaketen geschaffen werden, die der Anwender mit Parametern zu versorgen hat¹⁾.

2. Bedarf nach einfachem Methodenzugang besteht aber gerade bei solchen Anwendern, die sporadisch Analysen durchführen, die Methodenbasis selbst aber nicht verändern. Eigene Serviceabteilungen sind meist unwirtschaftlich und aus Akzeptanzgründen umstritten²⁾. Die typische betriebliche Situation dürfte eher so aussehen, daß der Problemspezialist aufgrund seiner Datenkenntnis Analysen selbst durchführt und beurteilt; dabei steuert er den Ablauf selbst, solange er seinem Wissensstand entspricht, und fordert Unterstützung an, sobald er sie benötigt. Ihm muß ein flexibles Dialogsystem zur Verfügung gestellt werden, das nahezu ohne DV-Kenntnisse bedient werden kann.

Entscheidend für den Erfolg einer individuellen Methodenbanklösung wird bereits die Vorauswahl des richti-

1) Vgl. z. B. die APL Econometric Planning Language EPLAN in Schober, F., Die Rolle der Datenverarbeitung in der Ökonometrie, in: Schober, F. und Plötzeneder, H. D. (Hrsg.), Ökonometrische Modelle und Systeme, München-Wien 1978, sowie die Vorgehensweise bei: Wolff, M.-R., Portfolio Management und Wertpapieranalyse vermittelt IDAMS, Heidelberg 1979.

2) Vgl. die entsprechende Diskussion über den Einsatz quantitativer Verfahren bei Böcker, F., Zur Akzeptanz formaler Entscheidungsmodelle im Marketing, Wirtschaftswissenschaftliches Studium 6 (1977), S. 305 ff.; Graham, R. J., Is Management Science arcane?, Interfaces 7 (1977), S. 63 ff.; Müller-Merbach, H., Tendenzen der Verwendung quantitativer Ansätze in der betriebswirtschaftlichen Forschung und Praxis, in: Müller-Merbach, H. (Hrsg.), Quantitative Ansätze in der Betriebswirtschaftslehre, München 1978, S. 11 ff., insbes. S. 17 ff.

gen Systemtyps für den jeweiligen Anwender sein. Methodenbankansätze der zweiten Kategorie für kompliziertere Verfahren existieren noch nicht, vielmehr hat man diese den Spezialisten überlassen¹⁾. Im folgenden soll deshalb speziell dieser Typ betrachtet werden.

1.3. Methodenspektrum von Methodenbanken

Empirische Untersuchungen in Unternehmen²⁾ und Hochschulen³⁾ zeigen das Übergewicht einfacher Verfahren bei praktischen Problemlösungen. Im Gegensatz dazu werden in der theoretischen Literatur zahlreiche Mängel dieser Verfahren und Ansätze zu ihrer Überwindung aufgezeigt. Häufig wird dabei die Verwendung multivariater Verfahren empfohlen; ihr Einsatz steigt ständig an.

Auch in Methodenbanken sind vorwiegend einfache Verfahren vertreten; neben ein- und zweidimensionalen Auszählungen sind dies vor allem Modelle für spezielle Aufgabenstellungen⁴⁾. Es gilt zu untersuchen, ob nicht gerade Methodenbanken das geeignete Instrument darstellen, den Einsatz anspruchsvoller Verfahren zu verbreiten bzw. die Methodenforscher zu unterstützen und zu überwachen.

¹⁾ Vgl. Krautter, J. und Scholz, P., Entwicklung ..., a.a.O.

²⁾ Vgl. Greenberg, B. A., Goldstucker, J. L., Bellinger, D. N., What Techniques are Used by Marketing Researchers in Business?, Journal of Marketing 41 (1977), S. 62 ff.

³⁾ Vgl. Leece, J. und Parker, F., Use and Misuse of SPSS, Software - Practice and Experience 8 (1978), S. 301 ff.

⁴⁾ Vgl. dazu das Marketing-Informationssystem der Henkel KGaA in Krautter, J. und Scholz, P., Entwicklung ..., a.a.O. sowie das System SABINE der Nielsen Company: A. C. Nielsen Company (Hrsg.), Nielsen Modell SABINE, interne Firmenschrift, Frankfurt o. J.

Für eine Methodengruppe ist im Kontext mit Methodenbanken zu prüfen:

- Besteht ein echter Bedarf für die Methode?¹⁾

Diese Frage ist von den Fach- und Methodenspezialisten gemeinsam zu lösen. Die Methodensammlung von Methodenbanken darf nicht mit Programmen belastet werden, die kein Anwender benötigt. Dabei ist zu bedenken, daß nur etwa 15 bis 30 % der in Fachzeitschriften veröffentlichten Modelle in Unternehmen implementiert sind, von denen wiederum die Mehrzahl nur einmal, nicht aber kontinuierlich angewandt wird²⁾.

- Kann die Methodenbank Zugangsbeschränkungen zu der Methode aufheben oder lockern?

Diese Frage ist von den Methoden- und Methodenbankspezialisten zu lösen. Es ist zu beobachten, daß häufig die - oft zufällige - Verfügbarkeit eines Verfahrens seinen Einsatz bestimmt. Durch erhöhte Zugangsmöglichkeiten, die zudem mit dem methodenbanküblichen Komfort verbunden sind, können solche Zugangsbeschränkungen prinzipiell abgebaut werden.

In den nächsten Jahren werden Hochschulabsolventen in der Praxis tätig sein, die multivariate Verfahren vom Studium her kennen und ihren Einsatz im Unternehmen fördern können; es ist an der Zeit, diese Verfahren in einsetzungsfähige Methodenbanksysteme zu integrieren.

¹⁾ Greenberg e.a. schreibt über die Ursachen des schwachen Einsatzes anspruchsvoller Methoden: "It is possible that this reflects a lack of understanding, a lack of capability, or a lack of need for these more sophisticated techniques", Greenberg, B. A. e.a., a.a.O., S. 66.

²⁾ Vgl. Böcker, F., Modellbezogene Akzeptanzprobleme formaler Entscheidungsmodelle im Marketing, in: Müller-Merbach, H. (Hrsg.), Quantitative ..., a.a.O., S. 227 ff., insbes. S. 229.

1.4. Zielsetzung beim Einbau der Clusteranalyse in eine Methodenbank

Die Clusteranalyse hat sich in den letzten Jahre als Entscheidungshilfsmittel in den Wirtschaftswissenschaften in den Vordergrund geschoben. Dabei geht der Einsatzbereich weit über das erste Anwendungsgebiet innerhalb der Betriebswirtschaft, das Marketing¹⁾, hinaus²⁾. In den empirischen Untersuchungen von Köhler und Uebele³⁾ sowie von Greenberg⁴⁾ liegt der Anteil der Unternehmen, die die Clusteranalyse verwenden, bei 13,5 bzw. 18 %; eine Nachfrage nach dem Verfahren kann somit unterstellt werden.

Hauptproblem der Clusteranalyse ist ihre Benutzerunfreundlichkeit⁵⁾. Jährlich erscheint eine Fülle von Publikationen⁶⁾ über das Verfahren; viele von ihnen stellen neue oder modifizierte Algorithmen vor. Der Schritt zu "Standard-Prozeduren"⁷⁾ ist noch nicht abgeschlossen.

1) Vgl. dazu Sheth, J. N., The Multivariate Revolution in Marketing Research, Journal of Marketing 35 (1971), S. 29; Frank, R. E. und Green, P. E., Numerical Taxonomy in Marketing Research, Journal of Marketing Research 5 (1968), S. 83 ff.

2) Vgl. Späth, H. (Hrsg.), Fallstudien zur Clusteranalyse, München 1977; Leisch, E., Bestandsaufnahme zum Einsatz multivariater Verfahren bei wirtschaftlichen Entscheidungsprozessen, Diplomarbeit, Nürnberg 1977.

3) Vgl. Köhler, R. und Uebele, H., a.a.O., S. 143.

4) Vgl. Greenberg, B. A. e.a., a.a.O., S. 65.

5) Vgl. z. B. Steinhausen, D. und Langer, K., Clusteranalyse: Einführung in Methoden und Verfahren der automatischen Klassifikation, Berlin-New York 1977, S. 183.

6) Ebenda, S. 5; sie schreiben von mehreren hundert, Cormack von über 1.000 einschlägigen Arbeiten pro Jahr. Cormack, R. M., A Review of Classification, Journal Royal Stat. Soc., Series A, 134 (1971), S. 321 ff.

7) Heinzlbecker hält deshalb den Einbau der Clusteranalyse in eine Methodenbank nur für sinnvoll, wenn diese von erfahrenen Datenanalytikern benutzt wird; Heinzlbecker, K., a.a.O., S. 272 f.

Entsprechend werden Clusteranalysen meistens von Hochschulen, Forschungsinstituten oder Verlagen durchgeführt. In realisierten Methodenbanksystemen ist das Verfahren bisher nicht eingesetzt.

Andererseits zeichnet sich gerade die Clusteranalyse durch eine Reihe von Eigenschaften aus, die für einen Methodenbankeinsatz sprechen:

- Methodenvielfalt

Es gibt bis heute kein Klassifikationsverfahren, das allgemein akzeptiert ist und in zufriedenstellender Weise alle Probleme lösen kann und das man somit generell empfehlen könnte¹⁾. Durch die Möglichkeit zur unterschiedlichen Kombination der einzelnen Bausteine, aus denen sich eine Analyse zusammensetzt, wird dieses Problem noch verschärft.

Die Schwierigkeit liegt hauptsächlich bei der Auswahl der Klassifikationsmethoden, die zum Teil willkürlich erfolgt²⁾. Die Unterstützung bei der Auswahl der geeigneten Methode ist aber ein typisches Merkmal für eine Methodenbank.

- Subjektivität

Auf der technischen Seite wird der Klassifikationsprozeß durch die Verwendung von Algorithmen objektiviert, dennoch bleibt die "Subjektivität objektiver Methoden"³⁾ bestehen. Clusterverfahren zielen nicht auf ein globales Optimum ab; vielmehr greift der An-

¹⁾ Vgl. z. B. Vogel, F., Probleme und Verfahren der numerischen Klassifikation, Göttingen 1975, S. 38 f.; Steinhausen, D. und Langer, K., Clusteranalyse ..., a.a.O., S. 182 f.

²⁾ Vgl. dazu Forst, H., Zur Klassifizierung von Städten nach wirtschafts- und sozialstatistischen Strukturmerkmalen, Würzburg 1974, S. 2.

³⁾ Buttler, G., Methoden zur Abgrenzung regionaler Arbeitsmärkte, Köln 1974, S. 63.

wender gezielt in den Ablauf ein und bringt dabei seine subjektiven Zielvorstellungen, seine Erfahrung und sein persönliches Urteil ein¹⁾. Ein interaktives Methodenbanksystem muß diese Eingriffe ermöglichen und - z. B. durch geeignete Hilfeleistungen - unterstützen.

- Notwendigkeit der Interpretation

Die Clusteranalyse ist dem menschlichen Denkprozeß sehr ähnlich und deshalb wohl von allen multivariaten Verfahren intellektuell am einsichtigsten; dennoch reicht das rein technische Ergebnis zum Verständnis der Lösung in der Regel nicht aus. Es ist vielmehr erforderlich, das Ergebnis geeignet darzustellen, es mit kleinen "Trivial-Bausteinen" zu verdeutlichen und es - in Relation zu anderen Ergebnissen - einzuordnen. Dabei muß der Anwender die Freiheit haben, geeignete Bausteine auszuwählen; die Methodenbank muß gezielt Informationen aus der Analyse ableiten und den Benutzer entsprechend unterstützen.

- Methodenverknüpfung

Oft reicht die Anwendung einer Clusteranalyse allein nicht, um ausreichend Einsicht in die Problemsituation zu vermitteln. Man wird deshalb verschiedene Analysen durchführen, z. B. mit wechselnden Merkmalen oder mit verschiedenen Algorithmen, und steht dann vor der Aufgabe, die Ergebnisse zu vergleichen

¹⁾ Vogel, F., Subjektivitäten bei der Klassifikation von Einheiten, in: Brockhoff, K., Dinkelbach, W., Kall, P., Pressmar, D. B. und Spicher, K. (Hrsg.), Proceedings in Operations Research 7, Würzburg 1978, S. 105 ff.

und zusammenzufassen. Häufig ist es auch erforderlich, andere Verfahren vor oder nach der Clusteranalyse durchzuführen, etwa um erste Tendenzen festzustellen oder das Ergebnis näher zu untersuchen¹⁾. Methodenbanksysteme gewährleisten diese Verknüpfung, ohne daß Datenverarbeitungskenntnisse beim Anwender erforderlich sind.

- Weiterverarbeitung der Ergebnisse

Die Clusteranalyse ist häufig eine Vorstufe zu der eigentlichen Analyse, die sich auf einzelne Cluster bezieht. Beispielsweise kann man Produkte nach ihrem Saisonverlauf klassifizieren, um anschließend die Prognose für den aktuellen Absatzverlauf - aus Kostengründen - nur noch für den "typischen Vertreter" jeder Klasse durchzuführen²⁾. Da solche Verbindungen zwischen Verfahren kaum einzugrenzen sind, muß praktisch jedes zusätzliche Verfahren nach der Clusteranalyse auf klassifizierte Daten zugreifen können.

Die direkte Forderung nach dem Einsatz der Clusteranalyse in einer Methodenbank findet sich seit Jahren in der Literatur, hauptsächlich in Verbindung mit Marketing-Informationssystemen³⁾; konkrete Hinweise für eine Realisierung der Ansätze fehlen.

Die vorliegende Arbeit will ein Pilotprojekt verwirklichen und zur Diskussion stellen. Die eigentliche Akzeptanzprüfung der Verfahren und vor allem ihrer Bau-

¹⁾ Vgl. z. B. Dubes, R. und Jain, A. K., Models and Methods in Cluster Validity, in: IEEE (Hrsg.), Proceedings der Conference on Pattern Recognition and Image Processing, Chicago 1978.

²⁾ Vgl. Späth, H., Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion, München 1975, S. 79 ff.

³⁾ Vgl. z. B. Breitung, A., a.a.O.; Diller, H., Marketing-Informationssysteme, Wirtschaftswissenschaftliches Studium 5 (1976), S. 97 ff., insbes. S. 102.

steine zur Darstellung, Interpretation und vergleichenden Analyse der Ergebnisse muß von Benutzern vorgenommen werden, die das System aus der Sicht ihrer Anwendungsprobleme beurteilen; deshalb wird eine Erprobung von ANCA durch Diplomarbeiten ebenso angestrebt wie ein Akzeptanztest in der industriellen Praxis.

2. Ausgangsbasis

2.1. Programmsysteme zur Clusteranalyse

2.1.1. Isolierte Programme

Isolierte Programme zur Clusteranalyse finden sich vorwiegend in Fachzeitschriftenaufsätzen, wo sie einem breiten Fachpublikum zur Diskussion gestellt werden. Sie müssen vor einer Anwendung meist noch mit einem Rahmenprogramm versehen werden, das das Einlesen der Daten und weitere - zum großen Teil problem-spezifische - unterstützende Routinen beinhaltet.

Erste Zusammenstellungen von Algorithmen, die eine größere Resonanz gefunden haben, wurden von Anderberg¹⁾ und Hartigan²⁾ sowie im deutschsprachigen Raum von Späth³⁾ und Steinhausen/Langer⁴⁾ veröffentlicht. Auch hierfür gilt, daß die Algorithmen nicht sofort aufrufbar sind; die Autoren geben aber Hinweise für Rahmenprogramme⁵⁾ bzw. für notwendige Programmänderungen⁶⁾ an. Zielgruppe aller Autoren ist jeweils der FORTRAN-kundige Analytiker, der sein Programm selbst erstellt und damit arbeitet.

1) Anderberg, M. R., Cluster Analysis for Applications, New York 1973.

2) Hartigan, J. A., Clustering algorithms, New York 1975.

3) Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O.

4) Steinhausen, D. und Langer, K., Clusteranalyse ..., a.a.O.

5) Vgl. z. B. bei Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O., die Seiten 74 und 115.

6) Vgl. dazu auch den Hinweis von Steinhausen/Langer auf eine vom Benutzer zu schreibende Subroutine; Steinhausen, D. und Langer, K., Clusteranalyse ..., a.a.O., S. 108 f.

2.1.2. Spezielle Programmsysteme zur Clusteranalyse bei Verlagen und Instituten

Am häufigsten wird die Clusteranalyse zur Zeit wohl bei den Typologiestudien verschiedener Verlage¹⁾ eingesetzt. Dennoch ist relativ wenig über die dazu verwendeten Programme bekannt, weil diese kommerzialisiert sind und die Anbieter infolgedessen an einer Veröffentlichung von Details, die eine Nachahmung erlauben würden, wenig Interesse haben²⁾.

Eine Übersicht über die meisten der verwendeten Programmsysteme liefert Wendt³⁾. Er berücksichtigt Systeme von

- ALGORITHM, Frankfurt
- BUREAU WENDT, Hamburg
- CISI, Paris
- COMPAGNON, Stuttgart
- DELTA-MARKTFORSCHUNG, Konstanz
- GfK, Nürnberg
- INTERLOGIC, Bielefeld
- INFRATEST, München
- ISBA, Frankfurt
- ITEM, Frankfurt
- IVE, Hamburg
- MORASCH UND PARTNER, Heidelberg
- SCS, Hamburg

¹⁾ Eine Übersicht geben z. B. Braunschweig, E. (Hrsg.), Typologien, o. O. 1975 sowie Böhler, H., Der Beitrag von Konsumententypologien zur Marktsegmentierung, Die Betriebswirtschaft 37 (1977), S. 447 ff.

²⁾ Vgl. Wendt, F., Typologische Analyse Systeme in der deutschen Marktforschung, Hamburg 1975, S. 29.

³⁾ Ebenda, zur Übersicht vgl. S. 30 f.; Literaturhinweise für die einzelnen Systeme findet man auf S. 126 ff.

Wendt unternimmt dabei auch den Versuch, ihre Unterschiede und Einsatzbedingungen klar herauszuarbeiten.

Charakteristisch für diese Systeme ist, daß sie von Spezialisten für eine feste Aufgabenstellung verwendet werden. Ein weiteres Merkmal ist die große Anzahl von Objekten und Merkmalen, die in eine Analyse eingehen.

Dementsprechend handelt es sich bei diesen Systemen ausschließlich um Systeme der Stapelverarbeitung, bei denen großer Wert auf Effizienz der Verarbeitung gelegt wurde. Bis auf eine Ausnahme werden wegen der großen Objektzahlen nur partitionierende, keine hierarchischen Verfahren¹⁾ verwendet.

2.1.3. Generelle Programmsysteme zur Clusteranalyse

Ansätze zur Erstellung generell verwendbarer Programmsysteme²⁾ finden sich in Unternehmen und bevorzugt an Hochschulen und Forschungseinrichtungen³⁾. Dabei werden häufig die in Abschnitt 2.1.1. beschriebenen Algorithmensammlungen als Methodenbasis verwendet⁴⁾. Teilweise erweitern auch "Erfinder" von Algorithmen ihr Programm um einige Optionen und bieten dann Programmsysteme an.

¹⁾ Der Begriff "partitionierende Verfahren" wurde von Späth übernommen, die grobe Unterteilung in die beiden Methodengruppen findet sich aber generell in der Standardliteratur; vgl. z. B. Anderberg, M. R., a.a.O.; Bock, H. H., Automatische Klassifikation, Göttingen 1974; Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O.; Vogel, F., Probleme ..., a.a.O.

²⁾ Eine Übersicht hierzu gibt Bock, H. H., Programme zur Clusteranalyse, Statistical Software Newsletter 2 (1976), S. 72 ff.

³⁾ Ebenda, S. 79; vgl. hierzu die Liste der betreuenden Stellen.

⁴⁾ So weiß Späth, daß seine Algorithmen in den verschiedensten Programmsystemen "am Laufen" sind.

Die bekannten Programmsysteme unterscheiden sich sehr nach Methodenumfang und Verarbeitungsmöglichkeiten. So enthalten etwa AUKL¹⁾ und CLAN²⁾ jeweils nur ein - im Fall von AUKL sehr spezielles - Verfahren; häufig finden sich verschiedene hierarchische Verfahren in einem Paket, etwa bei YGRP³⁾. Der Grund für die relativ starke Beschränkung dürfte neben dem Aufwand bei der Systemerstellung⁴⁾ vor allem die Vertrautheit der Autoren mit bestimmten Verfahren sein, die sich besonders in der jeweiligen Bevorzugung hierarchischer oder partitionierender Verfahren zeigt.

Die umfangreichste Methodensammlung beinhaltet das System CLUSTAN⁵⁾, es gilt gemeinhin als Vorbild für weitere Entwicklungen⁶⁾.

Der Vorteil dieser Systeme liegt darin, daß Methodenverknüpfungen zwischen den einzelnen Bausteinen und die Erzeugung konkurrierender Lösungen für eine Datenmenge möglich sind. Vom Anwender werden keine Programmierkenntnisse mehr verlangt. Er arbeitet vielmehr mit einer Kommandosprache, die z. B. bei CLUSTAN der Betriebssystemsprache verwandt ist.

1) Vgl. Faber, E. und Nollau, W., Über ein Verfahren zur automatischen Klassifikation, Darmstadt 1969.

2) Vgl. Lades, R., Schläger, W., CLAN, ein EDV-Programm zur Clusteranalyse für die Rechenanlage CD 3300, Arbeitspapiere des Betriebswirtschaftlichen Instituts der Friedrich-Alexander-Universität Erlangen-Nürnberg, Heft 26, Nürnberg 1974.

3) Vgl. Forst, H. T. und Vogel, F., Y-Group: Ein Fortran-IV-Programm zur hierarchischen Klassifizierung von Merkmalsträgern und Merkmalen, unveröffentlichtes Manuskript, Köln 1972.

4) Tryon und Bailey schätzen den Aufwand für ihr recht spezielles System BC TRY auf zehn Jahre; vgl. Tryon, R. C. und Bailey, D. E., Cluster Analysis, New York 1970.

5) Vgl. Wishart, D., CLUSTAN 1C Release 1 User Manual, London 1975; Schäffer, J. B., Beschreibung und Benutzeranleitung des Wishartschen Clusteranalyse-Pakets, Bericht MD 97 der Gesellschaft für Strahlen- und Umweltforschung, München 1974.

6) Vgl. z. B. Bock, H. H., Programme ..., a.a.O., S. 72; Vogel, F., Probleme ..., a.a.O., S. 39.

Der Nachteil der betrachteten Systeme liegt in der Regel darin, daß außer genau spezifizierten Interpretationsroutinen¹⁾ dem Benutzer keine Hilfen angeboten werden, insbesondere fehlen Übergangsmöglichkeiten zu Standard-Programmsystemen der Datenanalyse.

2.1.4. Clusteranalyse in Standard-Programmsystemen

Der direkte Weg, einen Methodenverbund zu ermöglichen, wäre die Aufnahme der Clusteranalyse in ein Standard-Programmsystem zur Datenanalyse. Bei Durchsicht der verschiedenen Systeme stellt man fest, daß die Clusteranalyse nur selten integriert ist und meist nur mit einem einzigen, oft sehr speziellen Gruppierungsverfahren²⁾.

Das weltweit am weitesten verbreitete System SPSS beinhaltet die Clusteranalyse nicht. Auch BMD kennt keine Clusteranalyse; in das aus BMD weiterentwickelte BMDP wurden drei hierarchische Verfahren sowie ein spezielles Verfahren zur gleichzeitigen Analyse von Objekten und Merkmalen aufgenommen. OSIRIS bietet je ein hierarchisches und partitionierendes Verfahren an.

¹⁾ Die ausführlichste Variante zur Analyse von Klassifikationsergebnissen liefert das Programm YDIAG, das in der Regel an YGRP anschließt. Vgl. dazu Forst, H. T. und Vogel, F., Diagnose von Klassifikationsergebnissen, unveröffentlichtes Manuskript, Köln und Kiel 1978.

²⁾ Vgl. Bock, H. H., Programme ..., a.a.O., S. 72.

2.2. Die interaktive Methodenbank MADAS

2.2.1. Zielgruppe

Ausgangspunkt und Richtschnur für ANCA ist die Methodenbank MADAS, die als erster Schritt zu einem branchenorientierten Marktdatenbanksystem¹⁾ erstellt wurde. Da ein solches Informationssystem eine breite Zielgruppe ansprechen muß, wurde auf seine Nutzung durch DV- und Methoden-"Halblaien" besonderer Wert gelegt. Dies ist von genereller Bedeutung, weil die Auswertung ähnlicher Datenbanken tendenziell immer mehr vom Fach- und Problemspezialisten, also nicht vom Methodenspezialisten, durchgeführt werden muß, der zudem mit mehreren Abfrage- und Analysesystemen arbeitet.

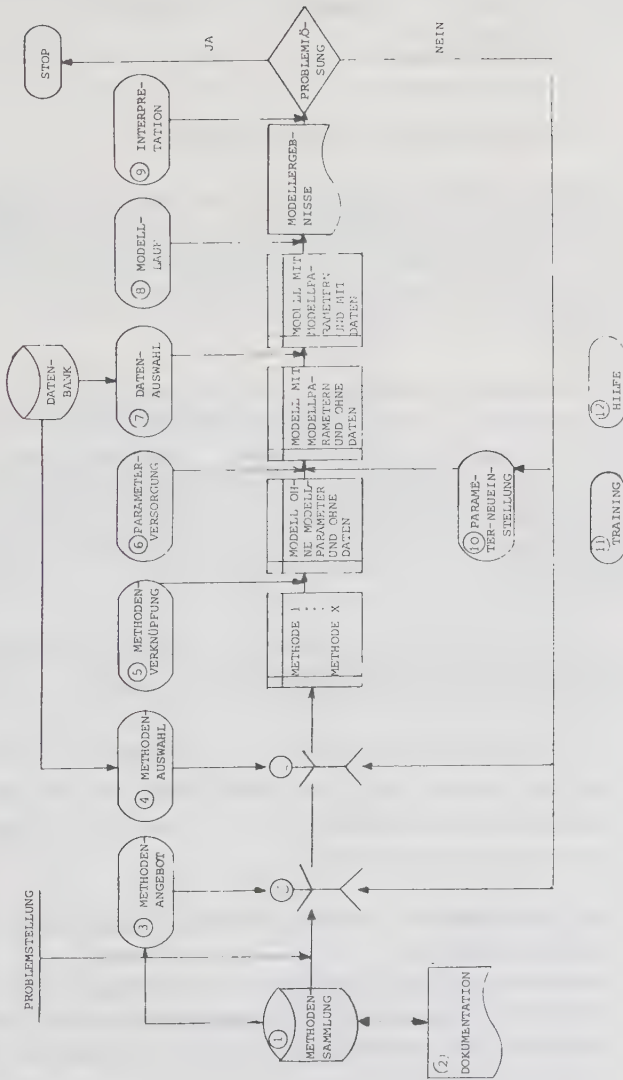
Vom Benutzer werden bei MADAS Grundkenntnisse der einzelnen Verfahren erwartet, die er anwenden will. Zur Überbrückung von Erinnerungslücken werden Hilfen bereitgestellt, die aber nicht bis zum computerunterstützten Unterricht reichen. Auf dem Weg von der Problemstellung zur Problemlösung kann der Analytiker Unterstützung in Anspruch nehmen, sei es bei der Auswahl der Merkmale und Verfahren oder bei der Interpretation der Ergebnisse. Dabei wird unterschieden zwischen solchen Maßnahmen, die fest in den Ablauf des Systems eingebettet sind, und solchen, die nur auf Wunsch abgerufen werden. Die Unterstützung, die eine "ideale" Methodenbank liefern kann, ist in den Ovalen der Abbildung 2.2.1./1 dargestellt²⁾; bei MADAS

1) Mertens, P., Möglichkeiten und Grenzen eines branchenorientierten Marktdatenbanksystems (MADAS), in: Hansen, H. R. (Hrsg.), Computergestützte Marketingplanung, München 1974, S. 181 ff.

2) Mertens, P. u.a., Verknüpfung von ..., a.a.O., S. 294.

2.2.1./1

Anforderungen an eine ideale Methodenbank



hat man sich bemüht, möglichst viele dieser Anforderungen zu verwirklichen. Sie sind auch Entwurfskriterien für ANCA und werden in den Abschnitten 3.1. und 3.2. aufgegriffen.

2.2.2. Verknüpfung mit der Datenbank

MADAS ist auf die Auswertung einer fest angeschlossenen Datenbank ausgerichtet; in der Pilotanwendung handelt es sich um Daten aus dem Haushaltspanel der GfK bzw. G&I. Eine Anpassung an andere Daten ist leicht möglich.

Die Verknüpfung von Daten- und Methodenbank ist mit zusätzlichem Aufwand verbunden. Struktur und Aufbau der Datenbank müssen dem System - unabhängig von einer möglichen Anwendung - vollständig bekannt gemacht werden; bestimmte Programmteile müssen generalisiert werden, was zu Effizienzverlusten bei Entwicklung und Ablauf führt.

Die enge Koppelung von Daten- und Methodenbank ist vor allem durch zwei Aspekte zu begründen:

- Eine Methodenanwendung soll "aus einem Guß" sein.

Von der Auswahl der relevanten Daten bis zur Ergebnisinterpretation darf es nicht notwendig werden, das System zu verlassen.

- Der Methodenbank bekannte Dateneigenschaften können genutzt werden.

Bezeichnungen und charakteristische Eigenschaften der Daten können in der Methodenanwendung zur Unterstützung der Methodenwahl und zur Erhöhung der Akzeptanz Verwendung finden.

Wirtschaftlich sinnvoll ist die Koppelung, wenn Daten über längere Zeit in gleicher Weise erfaßt und ausgewertet werden, unter Umständen sogar mit verschiedenen Paketen. Diese Bedingung liegt z. B. bei periodisch erhobenen Paneldaten vor, die mit mehreren Methoden analysiert werden.

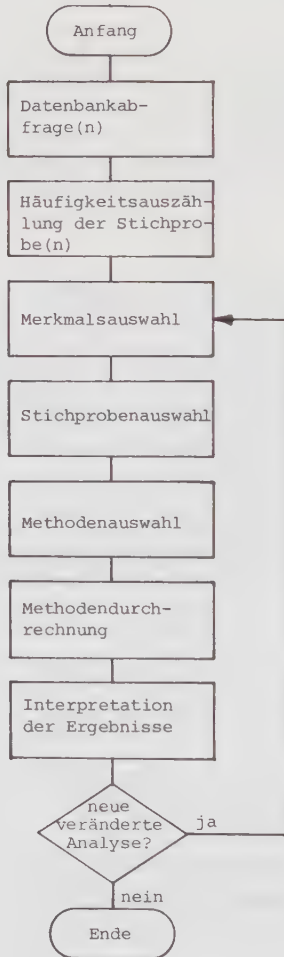
Abbildung 2.2.2./1 zeigt die einzelnen Phasen, die bei der Methodenanwendung in MADAS durchlaufen werden. Eine Anpassung an neue Daten wirkt sich durch die Verwendung genormter Schnittstellen nur in den Bausteinen "Datenbankabfrage" und "Häufigkeitsauszählung" aus.

2.2.3. Methodenspektrum in MADAS

Das Methodenspektrum in MADAS (Abbildung 2.2.3./1) ist hinreichend allgemein, um es unabhängig von einer speziellen Aufgabenstellung einsetzen zu können. Dennoch hat der ursprüngliche Anwendungsbereich, das Marketing, einen gewissen Einfluß speziell im Bereich der Modellanalyse genommen. Für die Auswertung der Marktdaten wurden ein Modell zur Berechnung von Marktanteilen und ein Modell zur Analyse von Wanderungsbewegungen implementiert.

2.2.2./1

Methodenanwendung in MADAS



2.2.3./1	Methodengruppen in MADAS
STRUKTURANALYSE	
- das sind Methoden der deskriptiven Statistik (univariate Strukturanalyse) - das sind Methoden der Korrelationsstatistik (bivariate Strukturanalyse) - das sind Verfahren aus dem Bereich der multivariaten Statistik (multivariate Strukturanalyse)	
STRUKTURVERGLEICH	
- das sind Methoden der Inferenzstatistik (Testverfahren) - das sind Methoden, die arithmetische Vergleiche erlauben	
ZEITREIHENANALYSE	
ABHÄNGIGKEITSANALYSE	
- das sind Methoden zur Ermittlung einseitiger Abhängigkeiten	
MODELLANALYSE	
- das sind komplexe Analyseverfahren, die der Benutzer aus den oben dargestellten Methoden und zusätzlichen Verfahren selbst entwickelt oder zukauf und adaptiert	

2.2.4. Verarbeitungsgrenzen

Für Methodenprogramme von MADAS mußten klar definierte Verarbeitungsgrenzen festgelegt werden; dabei spielt die Zahl "sieben" eine wichtige Rolle¹⁾: In einer Ana-

¹⁾ Dies entspricht der Faustregel, daß der Mensch nicht viel mehr als sieben Symbole gleichzeitig in seinem Gedächtnis speichern kann; vgl. hierzu Mertens, P., Die technische Gestaltung von Bildschirm-Dialogen, Online 15 (1977), S. 670 ff., hier S. 672.

lyse können maximal sieben Merkmale verarbeitet und ebenso viele Gruppen gebildet oder verglichen werden. Die Zahl der Objekte ist dagegen nur teilweise beschränkt; lediglich wenn eine Verarbeitung mit Hilfe von Matrizen erforderlich wird, dürfen nicht mehr als 100 Datensätze gleichzeitig analysiert werden.

Die Ursache der Einschränkungen sind zum einen die erwähnten Erfordernisse an Speicherplatz bei der Arbeit mit großen Matrizen; wenn die Beziehungen aller Objekte zueinander betrachtet werden sollen, verhält sich der notwendige Speicherplatz proportional zum Quadrat der Zahl der Objekte. Eine Auslagerung einzelner Teile der Matrix wäre denkbar, wegen ihrer Auswirkungen auf die Verarbeitungszeiten wird davon aber vorerst Abstand genommen. Zum anderen schränkt auch die Kapazität der zur Verfügung stehenden Bildschirmgeräte die Zahl der verarbeitbaren Elemente ein, weil gerade die Arbeit mit Methodenbanken die Darstellung der Ergebnisse - zur Zeit der Analyse - in übersichtlicher Form verlangt.

2.3. Methoden der Clusteranalyse in MADAS

2.3.1. Vorteile der Integration für MADAS

Die Integration einer Methodenbank mit einer speziellen Verfahrensgruppe in ein bestehendes Methodensystem setzt voraus, daß auf beiden Seiten der Ertrag den erforderlichen Aufwand übersteigt.

Auf der Seite von MADAS sind vor allem folgende Aspekte bedeutsam:

- Erweiterung des Methodenspektrums

Vor allem im Bereich der multivariaten Verfahren ergibt sich eine sinnvolle Ergänzung des bestehenden Spektrums (vgl. Abbildung 2.3.1./1). Von besonderer Bedeutung ist dabei, daß erstmals multivariate Verfahren vorhanden sind, die kein metrisches Skalenniveau erfordern.

2.3.1./1	Multivariate Verfahren in MADAS
Faktorenanalyse Multiple Korrelationsanalyse Partielle Korrelationsanalyse Kanonische Korrelationsanalyse Clusteranalyse Multivariate Varianzanalyse Multiple Kovarianzanalyse Multiple Diskriminanzanalyse Faktorielle Diskriminanzanalyse Berechnung von Klassifikationswahrscheinlichkeiten (Klassifikation)	

- Möglichkeiten zum Methodenverbund

Das System braucht nicht verlassen zu werden, wenn - z. B. vor bivariaten oder diskriminanzanalytischen Untersuchungen - eine erste Analyse der Daten mit der Clusteranalyse durchgeführt wird.

- Erweiterung der Analysemöglichkeit ausgewählter Datenmengen

In der Abfragesprache von MADAS müssen die Bedingungen fest vorgegeben sein; beispielsweise sollen alle

Hausfrauen ausgewählt werden, die über 4.000 DM Haushaltseinkommen haben, über 40 Jahre alt sind und im letzten Jahr eine Damenbluse gekauft haben. Häufig sind nur die Kriterien bekannt, nach denen selektiert werden soll, nicht aber die genauen Grenzen zur Differenzierung der Sätze; in einem analogen Beispiel würde man Hausfrauen suchen, die sich außer durch den Kauf einer Damenbluse durch hohes Einkommen und mittleres Alter beschreiben lassen. In diesem Fall kann man die Gesamtmenge mit Verfahren der Clusteranalyse strukturieren - wobei nach Art des Decision Calculus¹⁾ Gewichte für wichtige Merkmale vergeben werden können -, interessante Cluster auswählen und ebenso wie solche Gruppen untersuchen, die mit der Abfragesprache bestimmt wurden.

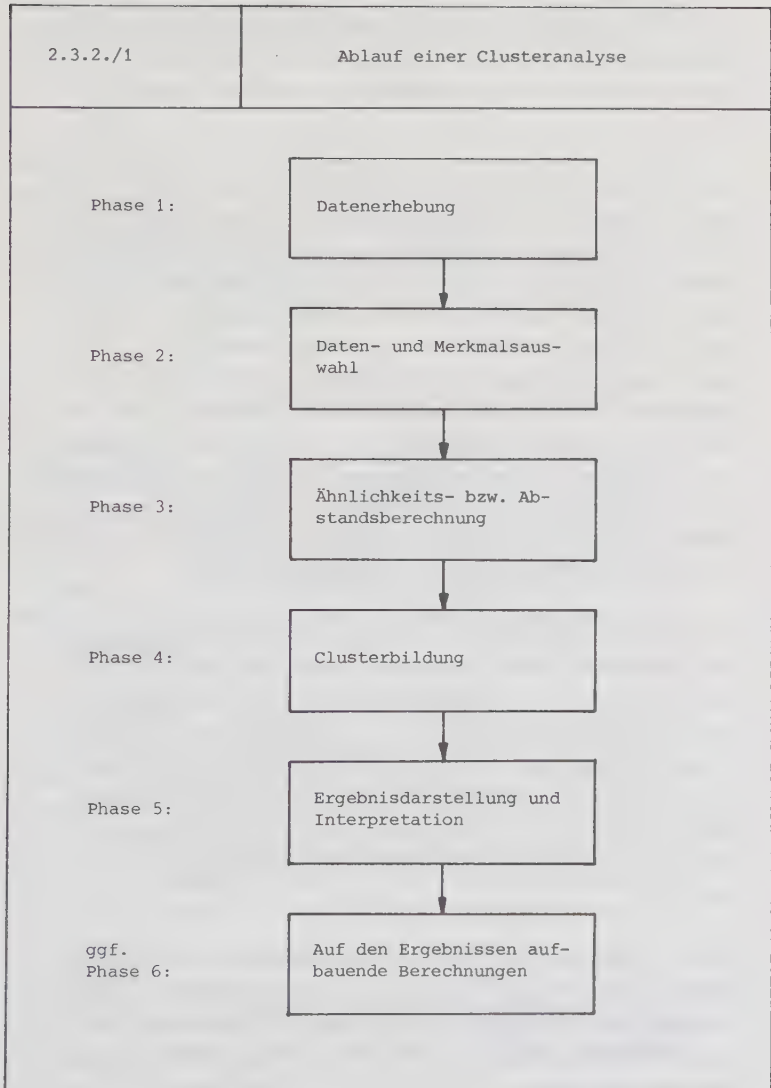
2.3.2. Vorteile der Integration für die Clusteranalyse

Clusteranalyse läuft in einem Prozeß ab, der neben dem reinen Algorithmus weitere Bausteine umfaßt. Diese entsprechen zum Teil generellen Funktionen von Datenanalysesystemen; aus Gründen der Wirtschaftlichkeit und der besseren Gewöhnung ist es sinnvoll, vorhandene Bausteine soweit als möglich zu übernehmen.

Anhand der Phasendarstellung der Abbildung 2.3.2./1 soll die Servicefunktion aufgezeigt werden, die ein System wie MADAS für die Clusteranalyse bieten kann²⁾.

1) Little, J. D. C., Models and Managers: The Concept of a Decision Calculus, Management Science 16 (1970), S. B-466 ff.

2) Die folgenden Punkte entsprechen im wesentlichen der "Wunschliste" von Bock an neue Programmsysteme zur Clusteranalyse; vgl. Bock, H. H., Programme ..., a.a.O.



Clusteranalyse beginnt mit der Erhebung und Verwaltung der relevanten Datenmenge (Phase 1). Diese Funktion übernimmt MADAS zentral für alle Methoden; eigene Routinen zur Datenverwaltung entfallen somit in ANCA.

In der Regel soll nicht die gesamte Datenmenge mit allen Merkmalen untersucht werden. Zur Definition und Erstellung von Stichproben existiert ein eigener Baustein in MADAS; zur Unterstützung der Merkmalsauswahl stellt MADAS Häufigkeitsverteilungen bereit, die Einblick in ihre Struktur vermitteln (Phase 2). In dieser Phase kann der Anwender auch Ausprägungen zusammenfassen, sei es aus methodischen Gründen - etwa um binäre Merkmale zu erhalten und den TANIMOTO-Abstand zu verwenden - oder gemäß einer bestimmten Zielsetzung, beispielsweise um die einzelnen Bundesländer beim "Wohnort des Verwenders" in Marktforschungsuntersuchungen nach NIELSEN-Gebieten zusammenzufassen.

Nach der Clusteranalyse erfolgt die Interpretation der manchmal schwer verständlichen Ergebnisse (Phase 5). Auch hier ist es sinnvoll, auf Standardbausteine zurückzugreifen, z. B. die Häufigkeitsverteilung in einzelnen Clustern im Vergleich zur Verteilung in der Gesamtmenge zu betrachten; dabei stehen alle Merkmale zur Verfügung, also auch solche, die nicht an der Klassifikationsentscheidung beteiligt waren.

Noch vorteilhafter ist die Verbindung von ANCA mit MADAS, wenn die gefundenen Cluster als "Stichproben" benannt und selbst zum Gegenstand von Analysen gemacht werden (Phase 6). In diesem Fall steht die ge-

samte Methodensammlung zur Verfügung, um ohne Wechsel des Systems sofort auf die neuen Daten angewandt zu werden.

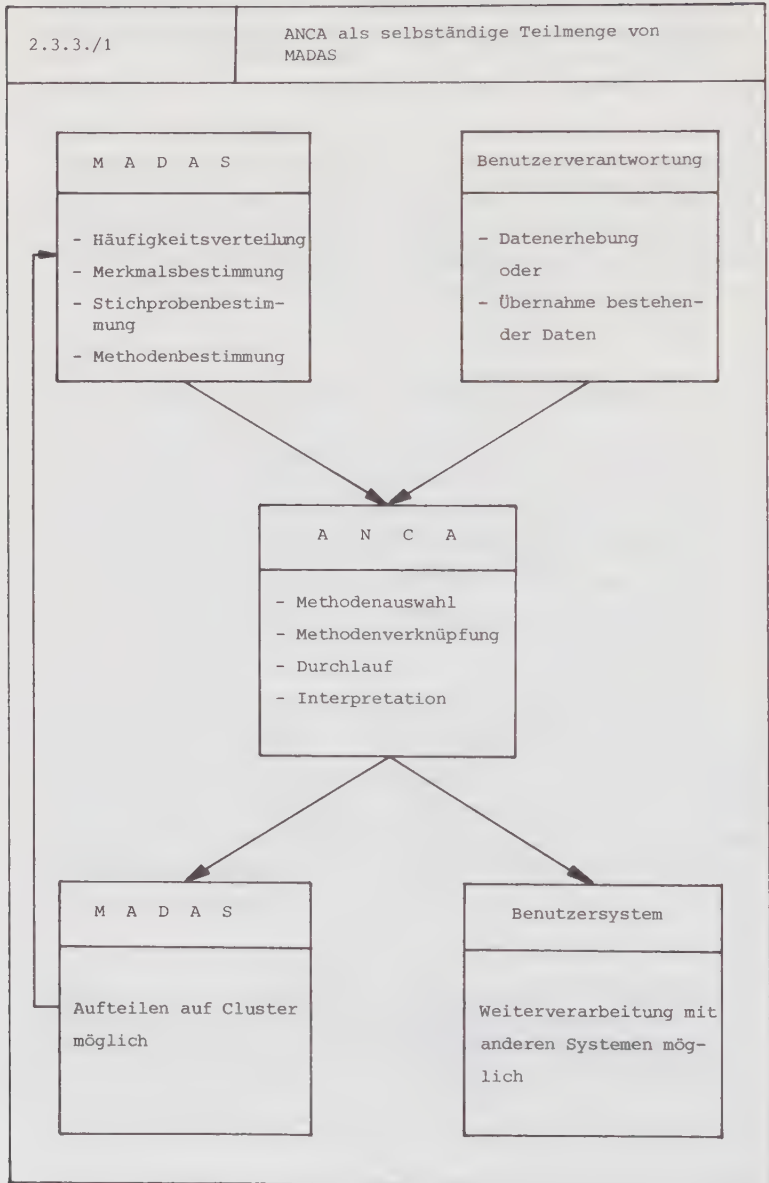
Die Verbindung bietet auch ausreichend Spielraum für Variationen des zugrundeliegenden Modells, der Methode und der Daten¹⁾, wenn Sensitivitätsanalysen durchgeführt werden sollen. Die Phasen drei und vier können mehrfach hintereinander angesprochen werden; ein Vergleich der verschiedenen Ergebnisse ist möglich.

2.3.3. ANCA als selbständige Teilmenge von MADAS

Die erwähnten Vorteile auf beiden Seiten lassen eine Integration von ANCA in das System MADAS sinnvoll erscheinen. Andererseits ist es nicht für jede clusteranalytische Untersuchung erforderlich, den gesamten Komfort eines umfassenden Methodenbanksystems bereitzustellen, zumal die erwähnten Effizienzverluste zu bedenken sind.

Als Konsequenz empfiehlt sich, ANCA und MADAS als getrennte Systeme zu realisieren, die über eine feste Schnittstelle miteinander kommunizieren (vgl. Abbildung 2.3.3./1). Bei gezielten Analysen kann der Anwender mit einem Teilsystem arbeiten; er soll einen problemlosen Einstieg finden und muß nur Angaben machen, die er für seine Aufgabenstellung benötigt. Anderer-

¹⁾ Zum Modell-, Methoden- und Datenverbund vgl. Schmidt, R., Zur Verbindung von Modellen, Methoden und Daten bei Unternehmensplanung mit EDV, in: Müller-Merbach, H. (Hrsg.), Quantitative ..., a.a.O., S. 95 ff.



seits darf durch diese lose Koppelung kein Verlust an Benutzerfreundlichkeit entstehen; insbesondere muß der Übergang von MADAS zu ANCA und zurück so reibungslos vor sich gehen, daß der Anwender annehmen kann, er arbeite mit einem System.

3. Die Methodenbank ANCA aus der Sicht des Anwenders

3.1. Systembasis von ANCA

3.1.1. Methodensammlung

Bei der Beurteilung der Methodensammlung eines gegebenen Methodenbanksystems kann eine Fülle von Kriterien herangezogen werden. Zu den wichtigsten gehören der Reichtum des Inhalts, die Möglichkeit zur Erweiterung um neue Methoden sowie die Richtigkeit der Programme¹⁾; diese Merkmale werden der folgenden Darstellung zugrunde gelegt.

3.1.1.1. Reichtum des Inhalts

Es wäre leicht, für ANCA eine reichhaltige Methodensammlung zu schaffen, verstünde man darunter nur die Aufnahme möglichst vieler verschiedener Algorithmen. Die Literatur bietet eine nahezu unübersehbare Anzahl von Verfahren, die zum Großteil auch in Form von DV-Programmen existieren. Schwieriger wäre es bereits, die Aktualität einer solchen Sammlung zu sichern. Die geringe Vereinheitlichung der Bezeichnungen und die manchmal schlechte Kontinuität bei der Entwicklung der Verfahren erschweren den Überblick²⁾. Das Hauptproblem liegt wohl darin, daß häufig ein auf eine Problemstellung speziell zugeschnittenes Programm verallgemeinert und zur "neuen" Methode erklärt wird.

¹⁾ Vgl. Mertens, P. und Bodendorf, F., Interaktiv nutzbare ..., a.a.O.

²⁾ Vgl. Hamacher und Preiser: "Die Vielzahl der bislang entwickelten Verfahren und der Mangel an objektiven Kriterien für ihren Einsatz macht es notwendig, die Eigenschaften der hier verwendeten Verfahren etwas näher zu erläutern"; Hamacher, B. und Preiser, K., Eine Infrastrukturtypologie am Beispiel des Landes Bremen, in: Späth, H. (Hrsg.), Fallstudien ..., a.a.O., S. 147 ff., hier S. 149.

Eine reichhaltige Methodenbank für unsere - in Abschnitt 2.2.1. definierte - Zielgruppe muß deshalb enger gefaßt werden. Es müssen "bewährte" Verfahren ausgewählt werden, die in der Literatur unbestritten sind und gewissermaßen den Weg "von der Forschung zur Normung" überstanden haben. Das Ergebnis mag eine für Methodenspezialisten "enttäuschende"¹⁾ Sammlung sein. Im Gegensatz zu einer "wissenschaftlichen Umgebung" kann aber ein hohes Angebot an Verfahren eher negative Effekte haben, wenn es nicht mit klaren, dem problemorientierten Benutzer vertrauten Einsatzbedingungen verbunden ist.

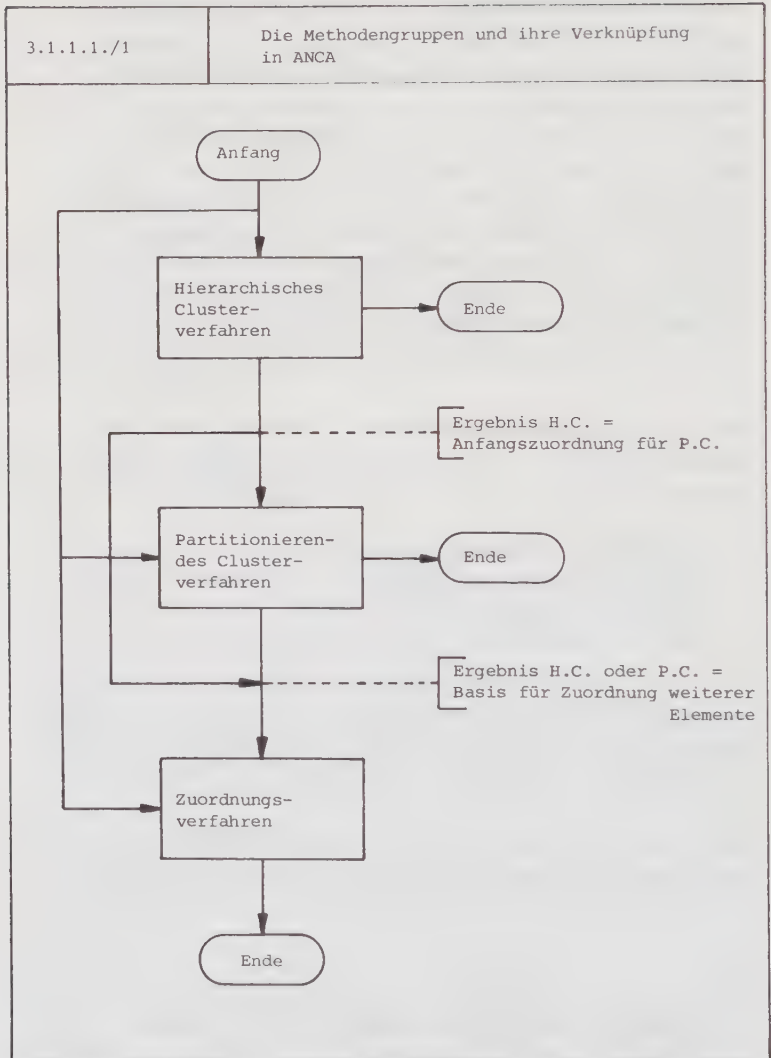
ANCA kennt deshalb in der ersten Ausbaustufe relativ wenige Verfahren. Sie werden zu drei Methodengruppen zusammengefaßt:

- Hierarchische Clusterverfahren
- Partitionierende Verfahren
- Zuordnungsverfahren

Wenn sich auch für oder gegen die einzelnen Verfahrensgruppen fast schon Schulen gebildet haben, scheinen alle drei erforderlich, zumal sie sich sinnvoll kombinieren lassen (vgl. Abbildung 3.1.1.1./1).

ANCA beinhaltet nur disjunkte exhaustive Verfahren, obwohl auch Argumente für eine Erweiterung um überlappen-

¹⁾ Vgl. z. B. den Diskussionsbeitrag von Bock in: Schaller, K., Eine Methodenbank zur Clusteranalyse (mit Diskussionsprotokoll), in: Gesellschaft für Klassifikation (Hrsg.), Klassifikation und Erkenntnis III, Studien zur Klassifikation, Band 6, Frankfurt 1979, S. 79 ff.



de¹⁾ bzw. nicht-exhaustive²⁾ Clusterverfahren sprechen. Im einzelnen sind zur eigentlichen Clustering folgende Verfahren enthalten (vgl. Abbildung 3.1.1.1./2):

3.1.1.1./2	Clusteralgorithmen in ANCA
7 hierarchische Verfahren³⁾: HIERCL 1 nach Johnson HIERCL 2 nach Johnson HIERCL 3 nach Sokal/Sneath HIERCL 4 nach Gower HIERCL 5 nach Sokal/Sneath HIERCL 6 nach Gower, Bock HIERCL 7 nach Ward das partitionierende Verfahren KMEANS⁴⁾ das partitionierende Verfahren CLUDIB⁵⁾ Zuordnung von Objekten ohne Veränderung des Schwerpunktes Zuordnung von Objekten mit Veränderung des Schwerpunktes	

¹⁾ Überlappende (nicht-disjunkte) Verfahren ordnen einige Objekte mehreren Clustern gleichzeitig zu. Vgl. z. B. das Verfahren KDEND in CLUSTAN. Allerdings bestehen für den durchschnittlichen Anwender große Verständnisprobleme wegen der Vielfalt der Ergebnisse in einem Lauf.

²⁾ Nicht-exhaustive Verfahren ordnen einzelne Objekte keinem der Cluster zu, sondern erklären sie zu Außenseitern. Allerdings ist das Problem der Abbruchkriterien bzw. Ähnlichkeitsschranken in keiner Weise gelöst.

³⁾ Die Verfahren wurden von Späth übernommen. Vgl. Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O., S. 170 ff.

⁴⁾ Vgl. ebenda, S. 68 ff.

⁵⁾ Vgl. Späth, H. (Hrsg.), Fallstudien ..., a.a.O., S. 68 ff.

Diese Verfahren werden für praktische Anwendungen im wirtschaftswissenschaftlichen Bereich am häufigsten eingesetzt und scheinen für einen "Härtetest" der Clusteranalyse am besten geeignet. Allerdings sind selbst von diesen Verfahren einige nicht nach außen hin einsichtig (HIERCL 4) bzw. nur in wenigen Fällen erfolgreich (HIERCL 1 wegen des Ketteffektes) und deshalb "auf Widerruf" in der Sammlung.

Verfahren, die auf der Analyse von Punktedichten¹⁾ beruhen, oder ähnliche - für Anwender nicht immer einsichtige - Verfahren fehlen völlig, auch wenn dann einige "kreative" Problemlösungen nicht nachvollzogen werden können; sobald sich hier "robuste" Standardverfahren herausgebildet haben, sollten sie aufgenommen werden.

Dagegen wurden bereits in der ersten Version Zuordnungsverfahren einbezogen. Neben der Möglichkeit, fast beliebig viele Objekte berücksichtigen zu können, bietet sich hier eine Möglichkeit, externe Fakten in den Analyseprozeß einzubringen: Der Anwender kann eine Aufteilung des Raumes nach seiner Vorstellung vorgeben und testen, ob diese durch die Zuordnung der realen Elemente erhalten bleibt oder verändert wird. Da die externe Zuordnung beispielsweise auch ein Ergebnis aus der Vorperiode sein kann, ist es möglich, die Stabilität einer Menge zu testen.

Für hierarchische Verfahren ist es erforderlich, entsprechende Maße für die Bestimmung der Ähnlichkeit

¹⁾ Vgl. hierzu Steinhausen, D. und Langer, K., Clusteranalyse ..., a.a.O., S. 156 ff. und die dort angegebene Literatur. Zur Kritik vgl. z. B. Wendt, F., a.a.O., S. 110.

bzw. Distanz von Objekten bereitzustellen. Kriterium für die Vollständigkeit bei ANCA ist das Abdecken aller Skalierungsvarianten¹⁾. Während für binäre und metrische Variablen bewährte Algorithmen existieren, treten bei nominalen und gemischten Daten Schwierigkeiten auf; ordinale Daten werden je nach Einschätzung des Analytikers als metrisch oder nominal behandelt. Im einzelnen sind die in Abbildung 3.1.1.1./3 enthaltenen Verfahren zur Bestimmung der Abstände implementiert.

3.1.1.1./1	Abstandsmaße in ANCA
- Euklidisches Distanzmaß - Quadriertes Euklidisches Distanzmaß - Tanimoto-Koeffizient - Übereinstimmungskoeffizient - Abstandsmaß MIX	

Für die einzelnen Maße sei auf die Literatur verwiesen²⁾. Das Verfahren MIX wurde in Anlehnung an Literaturbeispiele für ANCA erstellt; es wurde so konstruiert, daß neben der Verarbeitung gemischt-skaliertter Merkmale

¹⁾ Kompliziertere, z. B. hierarchische, Merkmalsstrukturen werden dabei nicht berücksichtigt. Vgl. hierzu Opitz, O., Neuere Entwicklungen zur Klassifikation mit qualitativen Daten, in: Gesellschaft für Klassifikation (Hrsg.), Klassifikation ..., a.a.O., S. 17 ff.

²⁾ Vgl. z. B. Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O., S. 14 ff; Vogel, F., Probleme ..., a.a.O., S. 78 ff.

auch die problemlose Verwendung von Merkmalen mit verschiedenen Spannweiten und verschiedener Anzahl von Ausprägungsstufen möglich ist¹⁾.

$$d_{ij} = \sum_{\substack{\text{binäre u.} \\ \text{nominale} \\ \text{Merkmale}}} \left\{ \begin{array}{l} 0 \text{ für } x_{ki} \neq x_{kj} \\ \frac{2}{a_k} \text{ für } x_{ki} = x_{kj} \end{array} \right. + \sum_{\substack{\text{ordinale u.} \\ \text{metrische} \\ \text{Merkmale}}} \frac{|x_{ki} - x_{kj}|}{\max_k - \min_k}$$

mit d_{ij} = Abstand zwischen dem i-ten und j-ten Element
 x_{ki} = Ausprägung des k-ten Merkmals beim i-ten Element
 a_k = Anzahl der Ausprägungsstufen des k-ten Merkmals
 $\min_k$ = kleinste Ausprägungsstufe des k-ten Merkmals
 $\max_k$ = größte Ausprägungsstufe des k-ten Merkmals

Bei partitionierenden Verfahren ist die vorgegebene Anfangszuordnung der Elemente bei einer festen Clusterzahl von großer Wichtigkeit für das Ergebnis; durch ihre Variation ist es möglich, ein Ergebnis auf seine Stabilität zu prüfen.

In ANCA sind deshalb - im Vergleich zu anderen Systemen - relativ viele Möglichkeiten der Anfangszuordnung implementiert (Abbildung 3.1.1.1./4).

¹⁾ Schaller, Klaus, Verfahren zur Darstellung und Analyse von Ergebnissen der Clusteranalyse - Entwurf, Implementation und Test an Daten der Marketingpraxis, Diplomarbeit, Nürnberg 1978, S. 23.

3.1.1.1./4	Angebot an Anfangszuordnungen für partitionierende Verfahren
Einer-Zuordnung: Startcluster: 1; Schaufelgröße: 1¹⁾ Zweier-Zuordnung: Startcluster: 1; Schaufelgröße: 2 Dreier-Zuordnung: Startcluster: 1; Schaufelgröße: 3 Zuordnung mit beliebigem Startcluster und beliebiger Schaufelgröße Zuordnung erzeugt durch Zufallszahlengenerator Zuordnung extern vorgegeben Zuordnung entspricht dem Endergebnis eines hierarchischen Verfahrens	

Besondere Probleme ergeben sich bei der Clusteranalyse häufig durch die schlechte Vergleichbarkeit der verwendeten Merkmale; dabei sind die unterschiedlich breiten Spannweiten der Merkmalsausprägungen und die unterschiedlich starken - meist subjektiv bestimmten - Gewichte der einzelnen Merkmale zu berücksichtigen.

ANCA kennt im einzelnen die in Abbildung 3.1.1.1./5 aufgeführten Verfahren zur Datenvorbehandlung.

¹⁾ Das Startcluster kennzeichnet dasjenige Cluster, dem das erste Element der Objektmenge zugeordnet wird. Die Schaufelgröße kennzeichnet die Zahl der Objekte, die sequentiell dem gleichen Cluster zugeordnet werden.

3.1.1.1./5	Angebot an Verfahren zur Datenvorbehandlung
Standardisierung von metrischen Daten ¹⁾ Transformation von ordinalen Daten ²⁾ Gewichtung von Merkmalen ³⁾ Maskierung von Merkmalen ⁴⁾ Umcodierung gemischter Daten in Binärform ⁵⁾ Ersetzen der Merkmale durch Faktoren ⁶⁾	

3.1.1.2. Erweiterbarkeit der Methodensammlung

Die vorgestellte Methodensammlung hat nur vorläufigen Charakter. Insbesondere für Anwender mit Vorkenntnissen auf dem Gebiet der Clusteranalyse ist eine Ergänzung der Methodensammlung um anspruchsvollere Verfahren wünschenswert. Darüber hinaus werden durch wissenschaftlichen Fortschritt weitere potentielle Verfahren einem größeren Anwenderkreis erschlossen.

¹⁾ Vgl. Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O., S. 19 f.; zur Problematik vgl. Vogel, F., Probleme ..., a.a.O., S. 10 ff.

²⁾ Vgl. Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O., S. 20.

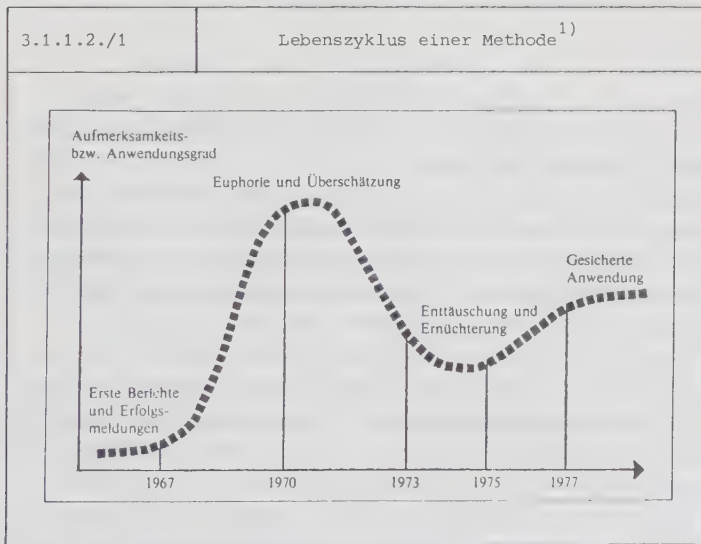
³⁾ Die Zweckmäßigkeit der Gewichtung ist umstritten; vgl. z. B. Morrison, D. G., Measurement Problems in Cluster Analysis, Management Science 13 (1967), S. B-779; Vogel, F., Probleme ..., a.a.O., S. 67 ff.

⁴⁾ Diese Variante ist ein Sonderfall der Gewichtung: Durch ein Gewicht von Null werden Merkmale von der aktiven Teilnahme an der Klassifikation ausgeschlossen. Vgl. z. B. Bock, Programme ..., a.a.O., S. 73.

⁵⁾ Vgl. zu dieser Variante der Behandlung gemischter Daten Anderberg, M.R., a.a.O., S. 30 ff. Wegen der generell ungelösten "besten" Behandlung gemischter Daten wurde noch das Abstandsmaß MIX aufgenommen (vgl. oben).

⁶⁾ Diese Variante ist bei Integration mit MADAS verfügbar (vgl. Abschnitt 3.4.). Die Methodenverknüpfung ist allerdings nicht unumstritten; vgl. z. B. Vogel, F., Probleme ..., a.a.O., S. 62 ff., insbes. S. 65.

ANCA ist von DV-technischer Seite her für Erweiterungen gerüstet (vgl. Abschnitt 4.2.2.). Aus problemorientierter Sicht müssen der Informationsbedarf und das Angebot an akzeptierten Verfahren berücksichtigt werden. Allerdings ist zu bedenken, daß die Implementierung eines neuen Verfahrens die Akzeptanz der gesamten Methodenbank gefährden kann. Es sollte daher in etwa das im "Lebenszyklus einer Methode" (vgl. Abbildung 3.1.1.2./1) beschriebene Reifestadium erreicht haben; im Zweifel sollte entweder ganz auf eine Erweiterung verzichtet oder ein Verweis auf eines der Standardpakete zur Clusteranalyse (vgl. Abschnitt 2.1.) gegeben werden.



¹⁾ Vgl. Geschka, H., Implementierungsprobleme bei der Anwendung von Ideenfindungsmethoden in der Praxis der Unternehmen, in: Pfohl, H.-C. und Rürup, B. (Hrsg.), Anwendungsprobleme moderner Planungs- und Entscheidungstechniken, Königstein 1978, S. 159 ff., insbes. S. 161. Der Lebenszyklus ist dort dargestellt am Beispiel der Ideenfindungsmethoden.

3.1.1.3. Richtigkeit der Methodensammlung

Hier sind vor allem drei Teilgesichtspunkte wichtig:

- Richtigkeit der zugrundeliegenden Methode
- Fehlerfreiheit des Programms
- ausreichende numerische Genauigkeit

Die Richtigkeit der Methode kann am ehesten durch die Übernahme von Standardverfahren aus der Literatur gewährleistet werden. Dabei ist allerdings Wert auf ein ausreichendes feed-back mit Anwendern zu legen; es empfiehlt sich außerdem, Algorithmen aus dem "Newsletter"-Teil theoretischer Zeitschriften erst nach gründlicher Prüfung zu übernehmen.

Methoden aus der Literatur sind unter Umständen öfter richtig als solche, die aus Standardpaketen stammen. Häufig werden solche Pakete von Spezialisten nicht ernst genommen und deshalb auch nicht geprüft, während "Laien" sich vielfach durch die in der Literatur beschworene "Computer-Gläubigkeit" auszeichnen und entweder unwahrscheinliche Ergebnisse mangels einschlägiger Methodenkenntnis ohne Zweifel akzeptieren oder die Methode per se für unbrauchbar halten.

Die Zahl der Fehler in den Programmen generell ist zum Teil noch erstaunlich hoch¹⁾. Besonderer Wert ist deshalb auf die Erstellung von Programmen zu legen, die leicht lesbar und modular aufgebaut sind; durch Routinen zur Beschleunigung der Laufzeit wird diese Zielsetzung manchmal wieder beeinträchtigt.

¹⁾ Pessemier, E. A., Data Quality in Marketing Information Systems, in: Farley, J. U. und Howard, J. A. (Hrsg.), Control of "Error" in Market Research Data, Lexington-Toronto-London 1975, S. 109 ff., hier S. 134.

ANCA verwendet überwiegend Methoden von Späth, die als Unterprogramme in das System integriert wurden. Die weite Verbreitung dieser Algorithmen garantiert ausreichende Rückmeldung von Fehlern und Problemen; die zugrundeliegenden Methoden gelten generell als bewährt und richtig.

Die erforderliche Genauigkeit hängt wesentlich von der Zahl der Iterationen ab¹⁾. Ferner ist zu prüfen, ob die Ergebnisse anschließend noch mit Verfahren bearbeitet werden, die ebenfalls Wert auf große Genauigkeit legen. Beispielsweise ist es bei Tests, die die Stabilität des Ergebnisses untersuchen, weniger wichtig, eine große Genauigkeit zu erreichen, als bei den Routinen zur Datenvorbehandlung, denen noch weitere Bausteine folgen²⁾.

3.1.2. Methodendokumentation

Für jede Methode in ANCA existiert eine Kurzbeschreibung, die eine Bildschirmseite nicht überschreitet. Dies wird für ausreichend erachtet, da "unmündige" Anwender nicht angesprochen werden sollen. Die Information kann jederzeit - sowohl in einer Informationsphase vor der eigentlichen Anwendung als auch bei Informationsmängeln während der Sitzung - abgerufen werden; zulässige Deskriptoren sind die Methodenbezeichnung oder ein Synonym.

¹⁾ Vgl. die Empfehlung von Späth, bei großer Objektzahl in KMEANS mit doppelter Genauigkeit zu rechnen. Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O., S. 89.

²⁾ Vgl. hierzu Forst, H. T., Zur Klassifizierung ..., a.a.O., S. 9.

Die vorhandenen Dokumentationsbilder sind zudem in einer Übersicht abrufbar, die ebenfalls eine Bildschirmseite nicht überschreitet.

Ein ausführliches Zugriffssystem, wie es etwa bei METHAPLAN realisiert ist, wurde für unwichtig erachtet; METHAPLAN ist aufgrund seiner Vielfalt (mehr als 300 Programme)¹⁾ mit ANCA nicht vergleichbar.

3.1.3. Interaktive Verarbeitung

Konventionelle Systeme zur Clusteranalyse laufen in Stapelverarbeitung ab. Diese Vorgehensweise wird meist mit den hohen Rechenzeiten bei Verarbeitung großer Datenmengen begründet. Im Gegensatz dazu ist bei ANCA interaktive Verarbeitung in Initialisierung und Ablauf gewählt worden. Im folgenden sollen die Vorteile dieses Vorgehens aufgezeigt und einige Nachteile angesprochen werden; eine Wertung erfolgt in Abschnitt 5.1.2.

Die Arbeit im Dialog bietet die Möglichkeit, eine Analyse mit wesentlich größerer Schnelligkeit durchzuführen, als es bei Batchverarbeitung möglich ist. Dabei kann zwar die Dialogsitzung selbst durchaus länger dauern als das Zusammenstellen der entsprechenden Steuerkarten und der Start der Verarbeitung im Stapelbetrieb, entscheidend ist jedoch die Gesamtzeit bis zum Erreichen der Ergebnisse. Durch die Dialogverarbeitung werden

¹⁾ Vgl. Esprester, A. C., Die Entwicklung ..., a.a.O.

- zeitaufwendige Fehlläufe verhindert, die durch falsche Parameterwahl hervorgerufen werden (bei der Batchverarbeitung stellt dies wegen der Vielzahl von Parametern eine häufige Fehlerursache dar);
- unbrauchbare Versuche rechtzeitig abgebrochen, z. B. bei ungleichgewichtigen Merkmalen;
- Wiederholungsläufe sofort gestartet.

Dialog ermöglicht individuelle Steuerung des Ablaufes. An vielen Stellen ist es wünschenswert bzw. erforderlich, daß der Anwender gezielt mit seinem Wissen in den Dialogablauf eingreift. Beispiele dafür sind die Festlegung einer Clusterzahl bei einem hierarchischen Verfahren, Bestimmung der Gewichte und Darstellung der Ergebnisse; für den unerfahrenen Benutzer sind vor allem Auswahl und Zusammenstellung der einzelnen Bausteine zur Clusteranalyse zu nennen. Entscheidungen, die sonst "blind" im voraus getroffen werden müßten, können in ANCA mit Empfehlungen unterstützt werden. Der Analytiker kann diese Empfehlungen annehmen oder ignorieren (wenn er Zusatzinformationen hat) bzw. weitere Informationen anfordern und dann entscheiden.

Deutlich im Vorteil ist der Dialog-Benutzer bei der Beurteilung des Ergebnisses. Bei Batchsystemen existieren im wesentlichen zwei Varianten, ein Ergebnis darzustellen:

- Beschränkung auf wenige Kenngrößen; der Analytiker muß selbst noch Nebenrechnungen und Zusammenstellungen vornehmen, um detaillierte Ergebnisübersichten zu erhalten
- Ausgabe aller denkbaren Varianten des Ergebnisses; der Anwender muß aus dem "Zahlen-Friedhof" mit Hilfe seines Zusatzwissens die Lösung erarbeiten.

Beide Versionen bauen eine hohe Akzeptanzhürde auf. ANCA kann durch die Dialogverarbeitung einen Mittelweg gehen:

Es stellt allgemeine Kennwerte zur Verfügung und fordert den Anwender zu einer Festlegung auf. Der somit definierte Weg wird vom System weiter verfolgt; daraus resultiert ein Ergebnis. Die einmal getroffene Festlegung kann selbstverständlich wieder rückgängig gemacht werden; daraus resultiert ein neues Ergebnis.

Ein Beispiel dafür ist die Clusterzahl bei hierarchischen Verfahren. In CLUSTAN werden alle denkbaren Kombinationen voll aufgeführt, bei überlappenden Verfahren sogar mit mehreren Lösungen für eine Clusterzahl. Für den "normalen" Anwender ist dies unbrauchbar; ANCA bietet deshalb die Entwicklung der Abstandssumme als Maß für die Qualität einer Lösung an und schlägt eine Clusterzahl vor. Nach der Entscheidung des Benutzers wird genau diese eine Lösung ausführlich erläutert. Ist das Ergebnis nicht zufriedenstellend, kann durch ein Kommando eine neue Clusterzahl bestimmt werden.

Besondere Vorteile bietet der Dialog aufgrund der hohen Akzeptanz graphischer Informationen bei Einsatz eines Graphik-Bildschirms (vgl. Abschnitt 4.1.3.). Die geschaffenen graphischen Interpretationsbilder können wahlweise auf dem Plotter oder dem Graphik-Display ausgegeben werden. Plotter-Zeichnungen werden offline erstellt und erfordern zum Teil beträchtlichen zeitlichen Aufwand; da ihre Verwendung immer beliebter wird, ist mit Warteschlangen und verzögerter Ausgabe zu rechnen. Im Dialog stehen Dendrogramme und Profildarstellungen der Cluster sofort zur Verfügung.

Sie können mit Standardfunktionen bearbeitet und verfeinert und bei Bedarf auf einer Hardcopyeinheit ausgedruckt bzw. zusätzlich am Plotter gezeichnet werden.

Bei interaktiver Verarbeitung kann eine Sensitivitätsanalyse im Sinne der gezielten Änderung von Parametern bei mehreren Läufen unterstützt werden. Dabei gibt der Ablauf einer Analyse Hinweise auf Veränderungen, die dann sofort auf ihre Wirksamkeit geprüft und gegebenenfalls revidiert werden können. Sinnvoll erscheint dies beispielsweise bei einer Variation der Gewichte für die einzelnen Merkmale, wenn ANCA mangelnde Trennschärfe einer wichtigen Variablen feststellt.

Mögliche Nachteile interaktiver Verarbeitung in ANCA können sein:

- umständliche Verarbeitung, speziell für erfahrene Benutzer
- lange Wartezeiten in einzelnen Dialogphasen
- Aufteilung einer Aussage auf eine Sequenz von Bildern

Die Verarbeitungszeiten eines gesamten Dialogs hängen im wesentlichen mit seiner Gestaltung zusammen. Bei der Konzeption von ANCA wurde darauf geachtet, daß außer reinen Ergebnisbildern keine Bilder ohne Auswahlentscheidung auftreten. An vielen Stellen ist es möglich, zukünftige Parameterentscheidungen gleich mit zu fällen und so den Ablauf zu beschleunigen; im Extremfall ist es möglich, durch eine Eingabe im Öffnungsbild alle Auswahlentscheidungen vorweg zu nehmen und bei den Interpretationsbildern wieder aufzusetzen.

Längere Wartezeiten treten beim Aufruf der eigentlichen Algorithmen auf. Die Rechenzeit ist dabei von

der Zahl der Objekte und der Zahl der Merkmale abhängig, wobei der Einfluß der beiden Faktoren je nach Verfahren variiert¹⁾. Es ist wichtig, durch Begrenzung dieser beiden Größen die Rechenzeit unter Kontrolle zu halten²⁾, so daß nach außen nur noch die Verwaltungszeit des Rechners in Erscheinung tritt, auf die das Programm keinen Einfluß hat. Spezielle "Number-crunchers" und mögliche Parallelverarbeitung lassen dabei eine Ausweitung der Grenzen in der nahen Zukunft erwarten³⁾.

Die Darstellungsmöglichkeiten im Dialog müssen von der normalen Kapazität alphanumerischer Bildschirmgeräte ausgehen, die 80 mal 20 Zeichen beträgt. Neben der Zahl der Elemente und der Merkmale wird nun auch die Zahl der Cluster kritisch, weil Übersichten über ein gesamtes Ergebnis möglich sein müssen.

Deshalb wurden in ANCA folgende Beschränkungen eingeführt bzw. von MADAS übernommen (vgl. Abschnitt 2.2.5.):

- Maximal 100 Objekte sind direkt verarbeitbar (zur Möglichkeit der Zuordnung vgl. Abschnitt 3.1.1.1.)
- Maximal sieben Merkmale können gleichzeitig verwendet werden
- Maximal zehn Cluster können gebildet werden

1) Vgl. hierzu Abschnitt 3.2.3.1. (Kostenschätzungen).

2) Besonders wichtig ist dies für Rechenzentren, die den Bedarf an Rechenzeit zwischen zwei Dialogeingaben restriktiv handhaben; beispielsweise kann dies heißen, daß jeweils weniger als eine CPU-Sekunde in Anspruch genommen wird. Eine Übertragung von ANCA auf ein mit IMS arbeitendes Rechenzentrum ist deshalb gescheitert.

3) Selbst das für lange Laufzeiten bekannte Verfahren AUKL soll dialogfähig gemacht werden; Nollau setzt dabei auf Parallelverarbeitung und den Einsatz von Arrayprozessoren. Vgl. Nollau, H. G., Ein Verfahren zur Automatischen Klassifikation, in: Gesellschaft für Klassifikation (Hrsg.), Klassifikation ..., a.a.O.

3.1.4. Training

Wenn auch aus den bekannten Programmsystemen zur Clusteranalyse keine besonderen Trainingsmoduln bekannt sind, steht doch die Notwendigkeit der Einübung in die verschiedenen Verfahren außer Zweifel.

Es bereitet keine Schwierigkeiten, ANCA mit einer speziellen Testdatei zu verknüpfen, mit der das Training gefördert werden kann. Die Daten sollten aus Gründen der Akzeptanz zwar für die Erarbeitung verschiedener Verfahren geeignet sein, in erster Linie aber aus dem gewohnten Anwendungsbereich des Planers stammen, um als realistisch anerkannt zu werden.

Für ANCA wurde kein spezieller Trainingsbaustein geschaffen; insbesondere ist es nicht möglich, während einer echten Analyse auf einen Trainingsmodus zurückzugreifen.

Dennoch ist das Training mit verschiedenen Verfahren möglich. Besondere Unterstützung bietet ANCA durch die Möglichkeit der Maskierung. So können Clusteranalysen mit wechselnden Merkmalskombinationen durchgeführt werden, ausgehend von gemischten Daten auch mit rein binären oder metrischen Daten. Im Extremfall ist es möglich, ein- und zweidimensionale Klassifikationen zu bilden, die leicht verständliche überschaubare Ergebnisse liefern. So läßt sich z. B. die Tendenz zu etwa gleich großen Clustern bei dem hierarchischen Verfahren nach Ward oder die Neigung zur Brückenbildung ("Ketteneffekt") der Methode single linkage (HIERCL 1) leicht zeigen.

3.2. Systemablauf bei einer Clusteranalyse

3.2.1. Daten- und Merkmalsauswahl

Im folgenden werden die einzelnen Phasen beschrieben, die ein Anwender bei einer Clusteranalyse mit ANCA durchläuft. Die Dialogbilder sind bei einer Analyse von Daten zur Verwendung von Kosmetika entstanden.

Die erste Phase einer Clusteranalyse umfaßt die Auswahl der Daten, die in einem Durchgang klassifiziert, und der Merkmale, die dazu verwendet werden sollen.

Bei Aufruf von ANCA werden dem Anwender die Zahl der verwendeten Datensätze und die Merkmale, die der Analyse zugrunde liegen, präsentiert (vgl. Abbildung 3.2.1./1).

3.2.1./1	Übersicht über Daten und Merkmale in ANCA

SIE HABEN SICH ZUR LOESUNG IHRER FRAGESTELLUNG FUER EIN VERFAHREN DER -->>CLUSTERANALYSE ENTSCIEDEN.	
SIE ARBEITEN MIT 100 SAETZEN UND FOLGENDEN 7 MERKMALEN PRO SATZ:	
LEBENSGEMEINSCHAFT	-->>SKALIERUNG: NOMINAL
ALTER	ORDINAL
EINKOMMEN	ORDINAL
ORTSGROESSE	ORDINAL
LEGE WERT AUF KOSMETIK	BINAER
ACHTE AUF MEINE GESUNDHEIT	BINAER
BIN GEFUEHLSBETONT	BINAER
- WEITERE INFORMATION UEBER DIE MERKMALE DURCH 'HILFE/MERKMALE'. - INFORMATION UEBER ALLE DIALOG-KOMMANDOS DURCH 'HILFE/KOMMANDOS'.	
SEND. \\>	

Dem System können diese Informationen aus zwei Quellen bekannt sein:

- Der Anwender hat sie vorher in einem Lauf mit dem System MADAS ausgewählt (vgl. Abschnitt 3.4.).
- Der Anwender hat durch Zuordnung einer Datendefinitionsdatei (vgl. Abschnitt 4.2.3.) seine Auswahl bekanntgegeben.

Der Durchlauf durch MADAS wird als Normalfall angesehen. Für diesen Fall kann auf eigene Routinen zur Datentransformation und -selektion verzichtet werden. Auch das Problem "fehlender Daten"¹⁾, das bei praktischen Analysen eine große Rolle spielt, aber nicht befriedigend gelöst ist, kann somit ausgeklammert werden.

Die Merkmalsauswahl hat nur vorläufigen Charakter; durch Rücksprung zu MADAS oder Änderung der Definitionsdatei kann sie vollständig revidiert werden, innerhalb von ANCA ist durch Maskierung eine Feinauswahl möglich. Im letzteren Fall werden die nicht ausgewählten Merkmale als "passiv" betrachtet und für Zwecke der Interpretation mit ausgewertet.

Die für unser Beispiel ausgewählten Haushalte werden durch soziodemographische Daten und Statements über ihre Einstellung zu Kosmetik, Gesundheit und Gefühlsbetonung beschrieben. Ziel ist es, eine Struktur der Haushalte zu finden, die konsistent erscheint; sie soll

¹⁾ Vgl. z. B. Wishart, D., Treatment of Missing Values in Cluster Analysis, in: Corsten, L. C. A. und Hermans, J. (Hrsg.), COMPSTAT 1978: Proceedings in Computational Statistics, 3. Symposium held in Leiden, Wien 1978, S. 281 ff.

anschließend mit der beobachteten Markenverwendung verglichen werden. Nähere Angaben über die Merkmale sind verfügbar und können über ein HILFE-Kommando (vgl. Abschnitt 3.2.8.) abgerufen werden.

3.2.2. Methodenangebot

Im nächsten Schritt wird die passende Methode für unsere Problemstellung gewählt. Der Suchprozeß dazu besteht aus einem Zusammenwirken von Methodenangeboten durch ANCA und Auswahlentscheidungen des Benutzers, der von ANCA unterstützt wird.

Dabei bietet ANCA Menus mit den jeweils zulässigen Verfahren für einen Teilschritt der Clusteranalyse an - z. B. alle zulässigen Verfahren zur Datenvorbehandlung - und fordert den Benutzer zur Entscheidung für eine oder mehrere der Methoden auf (vgl. Abbildung 3.2.2./1).

3.2.2./1	Methodenangebot in Auswahlmenüs
<pre> ***** IHRE DATEN SIND METRISCH BEVOR SIE EINE --->ABSTANDBERECHNUNG FUER DAS HIERARCHISCHE VERFAHREN DURCHFUEHREN ODER ZUM PARTITIONIERENDEN CLUSTERVERFAHREN VERZWEIGEN, HABEN SIE DIE MOEGLICHKEIT, IHRE DATEN AUFZUBEREITEN. (0) KEINE WEITERE DATENAUFBEREITUNG. (1) --->STANDARDISIERUNG (NUR BEI METRISCHEN U./O. ORDINALEN DATEN) (2) --->GEWICHTUNG BELIEBIGER MERKMALE (3) --->ZUSAMMENFASSUNG MEHRERER MERKMALE ZU FAKTOREN (WIE (1)) *PART VERZWEIGUNG ZUM PARTITIONIERENDEN CLUSTERVERFAHREN *INTERPRET INTERPRETATIONSHILF VON DER ENTSCHEIDUNG MIT ANWENDUNGSEMPFEHLUNGEN EINGABE: \> ***** </pre>	

Für Experten werden mögliche Alternativen aufgeführt; z. B. wird mit dem Kommando *PART der direkte Sprung in das partitionierende Verfahren ermöglicht, ohne das Menu zur Datenvorbehandlung zu beachten.

In der Literatur werden zwei Varianten genannt, die das Methodenangebot für eine Anwendung begrenzen:

- Differenzierung zwischen Normal- und Spezialmethoden¹⁾

Diese Variante geht davon aus, daß in eine Methodenbank alle Verfahren aufgenommen werden sollen, die mindestens von einem Anwender oder einer Anwendergruppe bedient werden können. Bei Beginn der Dialogsitzung muß sich jeder Benutzer identifizieren, wird einer Gruppe zugeteilt und bekommt nur das Angebot derjenigen Verfahren, deren Anwendung er beherrscht. Kompliziertere Verfahren oder solche, deren Ergebnisse schwer verständlich sind, bleiben Experten vorbehalten.

- Beschränkung des Methodenangebots aus Datenschutzmotiven²⁾

Die Notwendigkeit zum Datenschutz macht es erforderlich, Unbefugten den direkten Zugriff auf schützenswerte Daten zu verbieten. Eine Methodenanwendung kann dagegen erlaubt sein, wenn sie keine Rückschlüsse auf einzelne Daten zuläßt. Beispielsweise ist es im allgemeinen unbedenklich, den Mittelwert einer Ge-

¹⁾ Vgl. Krautter, J. und Scholz, P., Entwicklung ..., a.a.O.

²⁾ Hüber, R. und Lockemann, P. C., Information Protection by Method Base Systems, in: Yao, S. B. (Hrsg.), Proceedings der Fourth International Conference on Very Large Data Bases, Berlin 1978, S. 420 ff.

samtmenge errechnen zu lassen¹⁾. Eine Selektion nach logischen Kriterien und eine Darstellung der gefundenen Untermenge ist dagegen zu untersagen. Im Methodendstammsatz kann gespeichert werden, ob ein Verfahren generell, beschränkt auf die Gesamtmenge oder überhaupt nicht zugelassen ist. Damit kann sichergestellt werden, daß ein nicht-privilegierter Benutzer Methoden, die er nicht anwenden darf, gar nicht angeboten bekommt.

Beide Varianten sind in ANCA nicht realisiert; dem Anwender steht grundsätzlich das gesamte Angebot zur Verfügung.

Eine Differenzierung nach dem Schwierigkeitsgrad der Methoden scheint bei den derzeit angebotenen Verfahren nicht erforderlich; sie sind etwa gleich verständlich. Bei Ausbau der Methodensammlung ist dieser Punkt erneut zu überprüfen.

Datenschutzprobleme sind im Zusammenhang mit Clusteranalysen zu beachten, da das Verfahren eine gezielte Bildung von Untermengen erlaubt. Durch bewußte Manipulation kann - unter Umständen nach mehreren Experimenten - ein Ergebnis erzwungen werden, das den Schluß auf die gewünschte Information ermöglicht. Clusteranalyse mit sensiblen Daten sollte deshalb generell nur einem bestimmten Kreis zugänglich gemacht werden; eine Differenzierung zwischen einzelnen Verfahren ist nicht sinnvoll.

¹⁾ Die Einschränkung auf die Gesamtmenge ist erforderlich; ansonsten wäre es möglich, mittels Durchschnittsbildung zweier Teilmengen, die sich nur in einem Satz unterscheiden, auf diesen Satz zu schließen. Vgl. dazu Hüber, R. und Lockemann, P. C., Information Protection ..., a.a.O., S. 420 ff.

3.2.3. Methodenauswahl

3.2.3.1. Erscheinungsformen der Methodenauswahl

Bei der Konzeption einer "idealen" Methodenbank stehen Auswahlhilfen für den Anwender im Mittelpunkt. Dies gilt in besonderem Maße für ANCA, da die Clusteranalyse mit ihrer Methodenvielfalt solche Überlegungen geradezu herausfordert. Die Literatur bietet aber außer der Forderung nach Offenlegung der Methodenverwendung¹⁾ nur verstreute Hinweise auf Auswahlprinzipien; in den meisten Fällen bestimmt vielmehr die verfügbare Software den Methodeneinsatz²⁾. Eine umfassende anwenderorientierte Benutzerunterstützung, wie sie in Abschnitt 5.2.2. beschrieben wird, konnte deshalb nicht realisiert werden. Im folgenden werden nur die unmittelbar realisierten Auswahlhilfen³⁾ und Kriterien angesprochen; darüber hinaus gehende Überlegungen bleiben dem oben erwähnten Abschnitt 5.2.2. vorbehalten.

Eine Führung des Benutzers im Dialog ist realisiert. Dabei war unter zwei Alternativen zu entscheiden:

- Entwicklung einer Baumstruktur⁴⁾

Die Methoden sind im Speicher in Form einer Baumstruktur verknüpft; der Benutzer wird von der groben Zusammenfassung ("Clusteranalyse") bis zur ausführbaren Me-

¹⁾ So fordert Späth die Autoren der von ihm herausgegebenen Fallstudien zur Offenlegung der Methodenverwendung auf, weil Maßstäbe und Kriterien für die Methodenauswahl fehlen. Späth, H. (Hrsg.), Fallstudien ..., a.a.O., S. 7. In direkter Reaktion darauf verweist Schneider auf die Quintessenz der Überblicksliteratur, "daß jedes Verfahren seine Vor- und Nachteile hat": Schneider, W., Taxonomie der Gedächtnisleistungen schwacher und normaler Rechtschreiber, in: Späth, H. (Hrsg.), Fallstudien ..., a.a.O., S. 183.

²⁾ Vgl. ebenda, S. 7 und S. 137.

³⁾ Die Darstellung orientiert sich an der Einteilung der Auswahlhilfen bei Mertens, P. und Bodendorf, F., Interaktiv nutzbare ..., a.a.O.

⁴⁾ Vgl. Erbe, R. und Walch, G., An Interactive ..., a.a.O.

thode ("HIERCL 7 mit Euklidisch berechneten Abständen und Anwendung der Gewichtung") geführt (vgl. Abbildung 3.2.3.1./1).

- Zerlegung des Prozesses in einzelne Bausteine

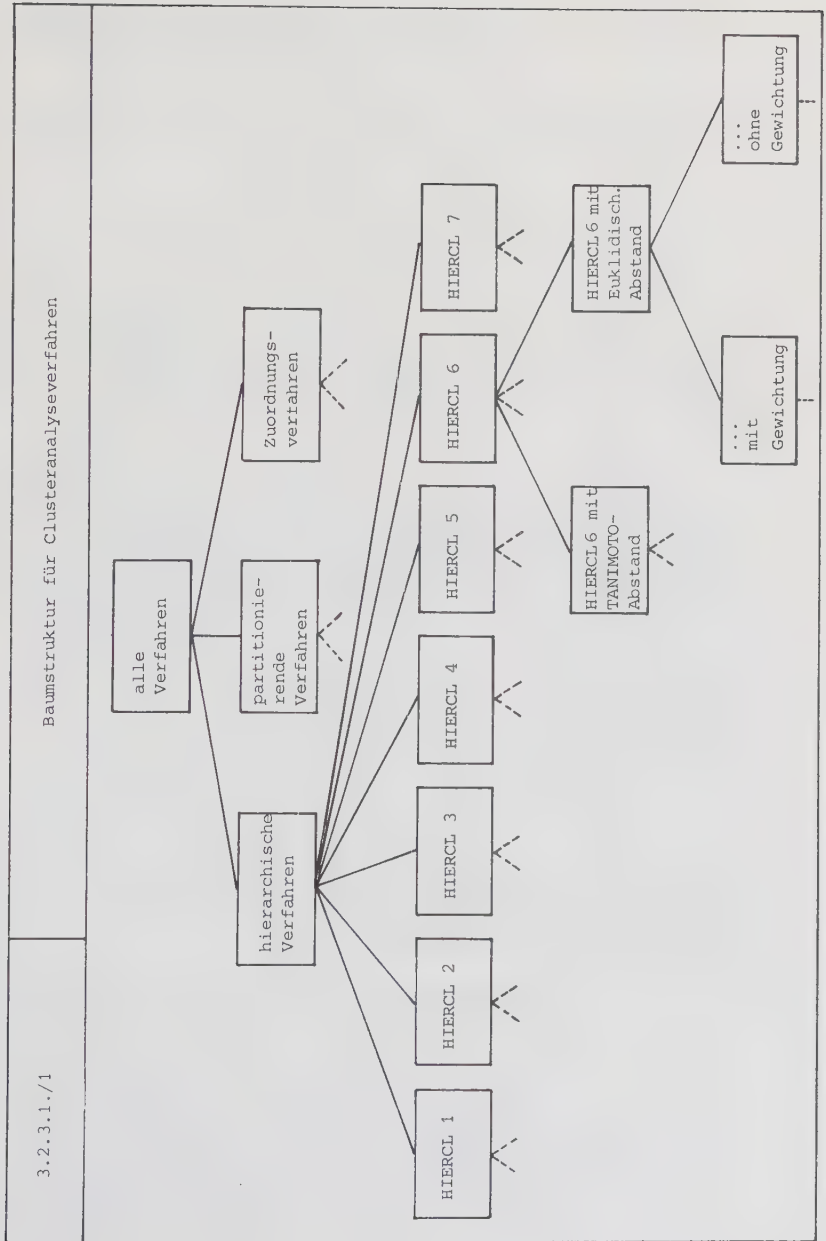
Die Clusteranalyse wird in ihre Bestandteile, z. B. Standardisierung, Bestimmung des Abstandsmaßes und des Clusteralgorithmus, zerlegt. Für jeden dieser Bausteine ist eine Auswahlentscheidung zu treffen; erst durch ihre Kombination wird die Analysemethode bestimmt.

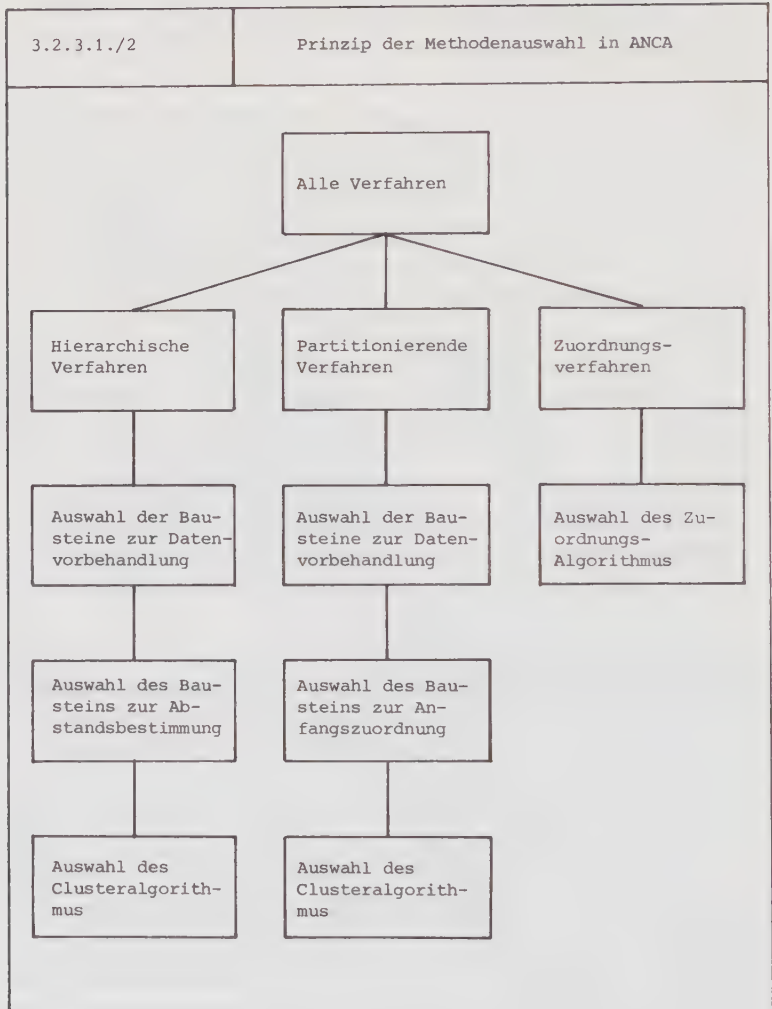
Von der Realisierung eines Baumes wurde abgesehen, weil aufgrund der unzureichenden Hinweise nur geringe Selektionsmöglichkeiten auf den einzelnen Stufen gegeben sind. Realisiert ist vielmehr nur die erste Stufe, d. h. die Auswahl der Methodengruppe durch den Anwender und die Auswahl des entsprechenden Zweiges durch das System. Alle folgenden Entscheidungen werden nach Variante 2 vorgenommen (vgl. Abbildung 3.2.3.1./2).

Für jeden einzelnen Baustein wurde geprüft, ob eine automatische Auswahl eines Verfahrens, die Reduktion der Vorgabemenge auf wenige Methoden oder nur Hilfen bei der benutzerindividuellen Entscheidung möglich sind. Es zeigt sich, daß eine automatische Auswahl über Entscheidungstabellen, wie sie bei den Korrelationsverfahren in MADAS realisiert wurde¹⁾, in keinem Fall möglich ist; dies wäre beim derzeitigen Stand der Forschung eine Einschränkung der Möglichkeiten speziell für den erfahrenen Analytiker, der sich häufig bewußt über allgemeine Richtlinien hinwegsetzt²⁾. Die eingebauten Selektionen und

¹⁾ Vgl. Mertens, P. u.a., Verknüpfung von ..., a.a.O., S. 324.

²⁾ Vgl. z. B. Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O., S. 51.





Empfehlungen sind in den Ablauf integriert, der in den folgenden Abschnitten dargestellt ist.

Verbote sind sinnvoll, wenn Verfahren aus theoretischen oder DV-technischen Gründen nicht anwendbar sind; Beispiele sind ein zu geringes Meßniveau der Merkmale oder ein zu hoher Bedarf an Internspeicher. Solche Verbote finden sich in den bekannten Programmsystemen zur Clusteranalyse recht selten; so unterscheidet CLUSTAN lediglich zwischen binären und numerischen Merkmalen, um einige Plausibilitätsprüfungen bei der Auswahl des Abstandsmaßes durchzuführen¹⁾; da ANCA von bekannten Skalierungen der Merkmale ausgeht, kann es relativ mehr Verfahren von vornherein ausschließen.

Kostenschätzungen als Auswahlhilfen sollen vor allem die leichtsinnige Vergeudung von Ressourcen verhindern helfen; sie finden sich beispielsweise in einem bei Makridakis u.a.²⁾ beschriebenen System. Auch ANCA könnte Unterschiede im zeitlichen Verhalten und somit bei den Kosten der Verfahren berücksichtigen; sie sind bei der Clusteranalyse wie folgt zu beobachten³⁾:

- Hierarchische Verfahren zeigen bei steigender Objektzahl einen überproportional anwachsenden Bedarf an Rechenzeit und Arbeitsspeicher, reagieren jedoch weniger auf eine Zunahme der Merkmalsanzahl.
- Partitionierende Verfahren sind sensibler in bezug auf die Zahl der Merkmale, weniger dagegen bei der Objektzahl.

In einer Testserie ergab sich folgendes Ergebnis (Abbildung 3.2.3.1./3):

¹⁾ Vgl. Wishart, D., CLUSTAN 1C ..., a.a.O.

²⁾ Vgl. Makridakis, S., Hodgson, A. und Wheelwright, S. C., An Interactive Forecasting System, Journal of the HP Educational Users Group 1 (1973) 2, S. 7 ff.

³⁾ Vgl. auch Vogel, F., Probleme ..., a.a.O., S. 230 f.

3.2.3.1./3	Zeitverbrauch für CLUSTAN-Jobs mit wechselnder Zahl von Merkmalen und Objekten ¹⁾		
Objekte	Variable	Part. Verfahren	Hier. Verfahren
100	7	16 sec.	17 sec.
200	7	31 sec.	101 sec.
300	7	45 sec.	274 sec.
100	33	29 sec.	20 sec.
100	50	35 sec.	25 sec.

Für ANCA war ein Einsatz von Kostenschätzungen bereits programmintern vorgesehen. Bei den derzeit in das System aufgenommenen neuen Verfahren sowie den maximal unterstützten 100 Objekten und sieben Merkmalen sind die Unterschiede aber vernachlässigbar²⁾. Diese Aussage gilt z. B. bei einer Einbeziehung des bereits angesprochenen Verfahrens KDEND nicht mehr; in der o. a. Untersuchung wurde für KDEND bereits bei 50 Objekten und sechs Merkmalen ein Zeitverbrauch von 203 Sekunden ermittelt³⁾. Da zudem die Ermittlung der Werte nicht unproblematisch ist, wurde ganz darauf verzichtet.

3.2.3.2. Auswahlhilfen für Verfahren der Datenvorbehandlung

Bei der Auswahl der Verfahren zur Datenvorbehandlung wird der Anwender informiert, welches Skalenniveau sei-

- 1) Die Tests wurden mit CLUSTAN durchgeführt, da ANCA nur 100 Objekte direkt verarbeitet. Die Zeiten gelten jeweils für den gesamten Job auf der CYBER 172 des Regionalen Rechenzentrums Erlangen. Vgl. Stumpf, H., Test und Vergleich der Programmsysteme ANCA und CLUSTAN zur Clusteranalyse, Studienarbeit, Nürnberg 1978.
- 2) Während die Verarbeitung von 100 Objekten und sieben Merkmalen bei einem Marktpreis von 1.200 DM/Std. etwa gleich viel kostet (DM 5,33 für 16 sec. bzw. DM 5,66 für 17 sec.), ist der Kostenunterschied zwischen beiden Verfahren bei 300 Objekten signifikant (DM 15,- für 45 sec., DM 91,33 für 274 sec.).
- 3) Stumpf, H., a.a.O., S. 42.

ne Daten haben (vgl. Abbildung 3.2.2./1, Zeile 1); da er mehr als eines der angebotenen Verfahren auswählen kann, wird er auch an vorab getroffene Entscheidungen erinnert (Abbildung 3.2.3.2./1).

3.2.3.2./1	Auswahlmenu zur Datenvorbehandlung nach Gewichtung
<pre> ***** IHRE DATEN SIND BINAER GEMISCHT UND WERDEN GEWICHTET MIT : 9 9 9 9 1 1 1 BEVOR SIE EINE --->ABSTANDRECHNUNG FUEER DAS HIERARCHISCHE VERFAHREN DURCHFUEHREN ODER ZUM PARTITIONIERENDEN CLUSTERVERFAHREN VERZWEIGEN, HABEN SIE DIE MOEGELICHKEIT, IHRE DATEN AUFZUBEREITEN. (0) KEINE WEITERE DATENAUFBEREITUNG. (1) --->GEWICHTUNG BELIEBIGER MERKMALE (2) --->UMCODIERUNG DER DATEN IN BINAERDATEN (FUEER GEMISCHTE DATEN) *PART VERZWEIGUNG ZUM PARTITIONIERENDEN CLUSTERVERFAHREN *INTERPRET INTERPRETATIONSBILD VOR DER ENTSCHEIDUNG MIT ANWENDUNGSEMPFEHLUNGEN EINGABE: \> ***** </pre>	

Keines der Verfahren könnte automatisch ausgewählt werden. Die Standardisierung ist nur bei metrischen und ordinalen Daten zulässig; sie ist sinnvoll, wenn die Skalenbreiten unterschiedlich sind. Allerdings finden sich nicht einmal in speziellen Aufsätzen zu diesem Thema konkrete Hinweise, wo die Notwendigkeit zur Standardisierung beginnt; es ist deshalb nur eine allgemeine Empfehlung möglich.

Die Gewichtung entspricht in der Regel subjektiven Vorstellungen des Benutzers; sie ist anzuwenden, wenn die

Merkmale von unterschiedlicher Bedeutung für die Zielsetzung des Analytikers sind. Eine Einschränkung auf ein oder mehrere Skalenniveau(s) ist nicht sinnvoll.

Die Ermittlung von Faktoren aus den gewählten Merkmalen wird besonders bei starken Korrelationen zwischen den Variablen empfohlen, ist aber sehr umstritten. Auch von Befürwortern werden keine genauen Einsatzbedingungen genannt. ANCA ermöglicht an dieser Stelle den Sprung in das Programm zur Faktorenanalyse, das an MADAS angekoppelt wurde. Der Anwender muß anschließend entscheiden, ob er mit den ursprünglichen Merkmalen oder den neu ermittelten Faktoren weiterrechnen will. Die momentan gängigen Verfahren der Faktorenanalyse verarbeiten nur metrische Daten; ANCA läßt sie deshalb nur bei diesem Skalenniveau und bei solchen Daten zu, die eigentlich ordinal sind, nach Ansicht des Anwenders aber als "quasi metrisch" gelten können.

Die Binärtransformation ist prinzipiell für alle Daten möglich, aber nur bei gemischten Daten sinnvoll.

Somit ergibt sich folgende sehr einfache Entscheidungstabelle (siehe Abbildung 3.2.3.2./2).

Unsere in Abschnitt 3.2.1. vorgestellten Daten sind binär gemischt. Entsprechend der Regel 5 stehen somit Verfahren der Gewichtung und der Binärtransformation zur Verfügung. Da wir wollen, daß die soziodemographischen Merkmale ein Übergewicht bei der Klassifikation erhalten, gewichten wir sie mit 9, die anderen Merkmale nur mit 1. Eine Binärtransformation nehmen wir nicht vor, weil uns das im nächsten Schritt empfohlene Maß MIX vorteilhafter erscheint.

3.2.3.2./2	Entscheidungstabelle zur Datenvorbehandlung				
Skalenniveau der Merkmale	1	2	3	4	5
Binär	J				
Nominal		J			
Ordinal			J		
Metrisch				J	
Gemischt					J
Empfohlene Verfahren					
Standardisierung			X	X	
Gewichtung	X	X	X	X	X
Faktorenanalyse			(X)	X	
Binärtransformation		(X)	(X)	(X)	X

3.2.3.3. Auswahlhilfen für Abstandsmaße

Auch bei den Abstandsmaßen ist keine automatische Auswahl eines Verfahrens möglich, sondern nur eine Vorselektion nach dem Skalenniveau, die unterschiedlich stark wirksam sein kann.

Der einfache und der quadrierte Euklidische Abstand sind definitionsgemäß nur bei metrischem Skalenniveau gültig. Aufgrund von praktischen Erfolgen, die sich bei Verwendung dieser Maße bei anderen Daten ergaben, wurden sie auch für ordinale Merkmale zugelassen. Der Unterschied zwischen beiden - größere Abweichungen bei einzelnen Merkmalen werden durch die Quadrierung deutlicher sichtbar - konnte nicht in Form einer Bedingung für die Entscheidungstabelle operational gemacht werden.

Der TANIMOTO-Koeffizient ist nur für binäre Daten definiert und für alle anderen Skalen verboten.

Der Übereinstimmungskoeffizient NOMIN verzichtet darauf, die Distanz zwischen einzelnen Ausprägungen zu ermitteln, sondern prüft nur auf Identität. Diese Vorgehensweise ist bei allen Skalen erlaubt, der hohe Informationsverlust läßt einen Einsatz bei metrischen Daten aber nicht sinnvoll erscheinen.

MIX wurde speziell für gemischte Daten geschaffen, ist aber auch bei stark unterschiedlicher Merkmalsstruktur sinnvoll. Es ist deshalb in jedem Fall im Angebot der Abstandsmaße enthalten.

Somit ergibt sich die - wiederum nur eindimensionale - Entscheidungstabelle der Abbildung 3.2.3.3./1:

3.2.3.3./1	Entscheidungstabelle zur Abstandsbestimmung				
Skalenniveau der Merkmale	1	2	3	4	5
Binär	J				
Nominal		J			
Ordinal			J		
Metrisch				J	
Gemischt					J
Empfohlene Verfahren					
Euklid			(X)	X	
Quadr. Euklid			(X)	X	
TANIMOTO	X				
NOMIN	(X)	X	X		X
MIX	(X)	X	X	X	X

Im Beispiel haben wir gemischte Daten, die nicht in binäre Merkmale transformiert wurden. Uns stehen deshalb gemäß Regel 5 zwei Maße zur Verfügung; da MIX versucht, die Eigenschaften der Ordinalskala bei drei der sieben Merkmale zu berücksichtigen, wählen wir es aus (vgl. Abbildung 3.2.3.3./2).

3.2.3.3./2	Auswahlmenu mit den Abstandsverfahren bei gemischten Daten
<pre> ***** IHRE DATEN SIND BINAER GEMISCHT UND WERDEN GEWICHTET MIT : 9 9 9 9 1 1 1 BESTIMMUNG EINES -->>DISTANZMASSES FUEH DIE ABSTANDSBERECHNUNG METHODEN DER ABSTANDSBERECHNUNG: (1) -->>NOMIN = UEBEREINSTIMMUNGSKOEFFIZIENT (FUEH ALLE DATEN) (2) -->>MIX (FUEH ALLE = SPEZIELL GEMISCHTE = DATEN) *PART VERZWEIGUNG ZUM PARTITIONIERENDEN VERFAHREN EINGABE: \> ***** </pre>	

3.2.3.4. Auswahlhilfen für Clusteralgorithmen

Zur eigentlichen Klassifikationsentscheidung stehen die drei Methodengruppen der Abbildung 3.2.3.4./1 zur Verfügung.

Der Anwender entscheidet je nach Zielsetzung; dabei ist wesentlich, ob er eine feste Zahl von Clustern im Sinn hat oder eine komplette Hierarchie der Daten erstellen will, ob er von vorhandenen Clustern ausgeht oder völlig neu klassifizieren möchte (vgl. Abbildung 3.2.3.4./2).

3.2.3.4./1	Auswahlmenu mit den Methodengruppen
***** ES STEHEN VERSCHIEDENE CLUSTER-ALGORITHMEN ZUR AUSWAHL: (1) HIERARCHISCHE VERFAHREN (2) PARTITIONIERENDE VERFAHREN (3) DIE ELEMENTE WERDEN EINGELESENEN CLUSTERSCHWERPUNKTEN ZUGEORDET FOLGENDE SYSTEMHILFE KANN IHNEN DIE ENTSCHEIDUNG ERLEICHTERN: (0) KURZBESCHREIBUNG BEIDER VERFAHREN FALLS SIE DIE NUMMER DES GEWUNSCHTEN HIERARCHISCHEN VERFAHRENS ODER DAZU DIE CLUSTERANZAHL SCHON WISSEN, KOENNEN SIE DURCH MEHRERE EINGABEN EINIGE BILDSCHIRME UEBERSPRINGEN. BEISPIELEINGABE: 1/5 HIERARCHISCHES VERFAHREN NR.5 1/3/10 HIERARCHISCHES VERFAHREN NR.3 MIT EINGABE: BILDUNG VON 10 CLUSTERN \>1 *****	

3.2.3.4./2	Auswahlkriterien für Methodengruppen							
	1	2	3	4	5	6	7	8
Clusterzahl bekannt	J	J	J	J	N	N	N	N
Hierarchie erwünscht	J	J	N	N	J	J	N	N
Clusterschwerpunkte vorgegeben	J	N	J	N	J	N	J	N
Hierarchisches Verfahren sinnvoll		X				X		
Partitionierendes Verfahren sinnvoll				X				
Zuordnungsverfahren sinnvoll			X					
Keine Variante sinnvoll bzw. widersprüchliche Angabe	X				X		X	X

Da wir eine erste Analyse durchführen und noch keine Vorstellung über eine sinnvolle Clusterzahl haben, entscheiden wir uns für hierarchische Verfahren.

Es stehen sieben Verfahren zur Wahl (vgl. Abbildung 3.2.3.4./3).

3.2.3.4./3	Auswahlmenu der hierarchischen Verfahren
------------	--

```

*****
ES STEHEN IHNEN 7 VERSCHIEDENE HIERARCHISCHE ALGORITHMEN ZUR VERFUEGUNG.

      EIGENSCHAFTEN, ANWENDUNGSEMPFEHLUNGEN
(1) H1:  INEINANDER VERSCHLUNGENE CLUSTER (VORSICHT: OFT KETTENBILDUNG)
(2) H2:  ERKENNUNG 'RUNDER' CLUSTER
(3) H3:  GUTE ERKENNUNG VON AUSSENSEITERN
(4) H4:  BILDUNG MEHRERER GROSSER CLUSTER
(5) H5:  GUTES VERFAHREN BEI VIELEN GLEICHEN ELEMENTEN
(6) H6:  GUTES VERFAHREN BEI RELATIV HOMOGENER ELEMENTMENGE
(7) H7:  BILDUNG GLEICH GROSSER CLUSTER. WIRD ALS 1. VERFAHREN EMPFOHLEN.
              (BESONDERS BEI METRISCHEN DATEN)
(0) ALLE VERFAHREN WERDEN NACHEINANDER DURCHGERECHNET.
      EMPFEHLUNG BEI UNBEKANNTER STRUKTUR DER DATENMENGE:
      DURCHFUEHRUNG VON H7 UND H3, VERGLEICH DER GESICHERTEN ERGEBNISSE.

FALLS SIE DIE GEWUNSCHTTE CLUSTERANZAHL SCHON WISSEN, KÖNNEN SIE DURCH ZWEI
EINGABEN EINIGE DIALOGSCHRITTE UEBERSPRINGEN.
      BEISPIELEINGABE:  3/10  HIERARCHISCHES VERFAHREN NR.3 MIT
      EINGABE:          BILDUNG VON 10 CLUSTER'1
      |>'
*****

```

Eine Entscheidungstabelle kann hier nicht angeboten werden; über die Einsatzbedingungen und -möglichkeiten der verschiedenen Verfahren wird von Methodenforschern noch diskutiert.

Als erstes Entscheidungskriterium kommt wieder das Skalenniveau in Frage. Zwar lassen sich einzelne Verfahren auf unterschiedliche Annahmen über das zugrunde liegende Skalenniveau der Merkmale zurückführen; in den Lehr-

büchern wird allerdings generell empfohlen, mehr als eines der Verfahren zu verwenden, ohne dabei auf die Skalierung Rücksicht zu nehmen. So bietet Späth die in ANCA übernommene Variante an, alle sieben Verfahren sequentiell durchzurechnen.

Ein weiteres Entscheidungskriterium ist die Zielsetzung der Analyse. Beispielsweise kann es sinnvoll sein, die Gesamtmenge in "Typen" aufzuteilen, die den Raum in etwa gleichmäßig aufteilen; dabei werden dünn besetzte Räume zu größeren Einheiten zusammengefaßt. In einem anderen Anwendungsfall sollen aus einer relativ einheitlichen Menge bewußt die "Störobjekte", z. B. unzufriedene Käufer, eliminiert und als Außenseiter genauer analysiert werden. Auch für dieses Kriterium lassen sich keine konsistenten Erkenntnisse finden.

Diese wenigen Hinweise und eigene Tests wurden nicht für ausreichend gehalten, um eine automatische Selektion vorzunehmen; sie erscheinen deshalb nur als Anwendungsempfehlungen. Neben Verweisen auf das Skalenniveau (HIERCL 7) und die Zielsetzung der Analyse (HIERCL 3) können auch negative Eigenschaften, beispielsweise der Ketteffekt von HIERCL 1, aufgenommen werden.

Für unsere gemischten Daten kann kein optimales Verfahren empfohlen werden; da HIERCL 7 in der Vergangenheit gute Ergebnisse bei vergleichbaren Daten gebracht hat, verwenden wir es auch in unserer Untersuchung.

3.2.3.5. Auswahlhilfen für partitionierende Verfahren

Bei partitionierenden Verfahren ist momentan nur das Standardverfahren KMEANS in der Methodensammlung; ei-

ne Auswahl erübrigt sich somit. Dagegen muß der Anwender unter den angebotenen Anfangszuordnungen wählen (vgl. Abbildung 3.2.3.5./1). Diese Entscheidung kann von ANCA nicht unterstützt werden. In der Literatur finden sich neben dem Hinweis, daß die Anfangszuordnung Einfluß auf das Ergebnis nimmt, nur der Vorschlag, einige Varianten durchzuspielen und die Ergebnisse zu vergleichen, sowie subjektive Präferenzen für eine zufällige oder gezielt ausgewählte Anfangszuordnung.

3.2.3.5./1	Auswahlmenu der Anfangszuordnungen
<pre> ***** IN DIESEM VERFAHREN FUER METRISCH SKALIERTE DATEN IST EINE SINNVOLLE ANFANGSZUORDNUNG DER ELEMENTE ENTSCHEIDEND FUER DIE CLUSTERBILDUNG. FOLGENDE MOEGlichkeiten STEHEN ZUR AUSWAHL: (1) -->>ABZAHLALGORITHMUS MIT 1ER-ZUORDNUNG (2) -->>ABZAHLALGORITHMUS MIT 2ER-ZUORDNUNG (3) -->>ABZAHLALGORITHMUS MIT 3ER-ZUORDNUNG (4) -->>ZUFALLIGE ANFANGSZUORDNUNG MIT HILFE EINES ZUFALLSZAHLENGENERATORS (RANDOM) (5) ZUORDNUNGSRoutine -->>CLAUF (ERGEBNISAUFBEREITUNG) (6) HIERAKHISCHES CLUSTERVERFAHREN (7) EXTERN VORGEGEBENE ANFANGSZUORDNUNG EINGABE: \>7 ***** </pre>	

3.2.4. Methodenverknüpfung

Die Verknüpfung von Methoden wird innerhalb von ANCA automatisch vorgenommen; darüber hinaus sind Verknüpfungen mit weiteren Methoden in MADAS möglich (vgl. Abschnitt 3.4.).

Die Verknüpfung der einzelnen Teilmethoden einer Clusteranalyse geschieht aufgrund der Angaben, die vom Benutzer

im Dialog gemacht werden. Die Verträglichkeit der einzelnen Teilmethoden wird geprüft; eventuell wird eine Verknüpfung verboten.

Abweichungen von der geplanten Vorgehensweise und Korrekturen bereits bestimmter Verfahrenskombinationen sind aufgrund der Dialogverarbeitung leicht möglich.

An dieser Stelle soll noch einmal auf die Verknüpfung eingegangen werden, die in Abbildung 3.1.1.1./1 dargestellt ist. Bei Unsicherheit über die Clusterzahl empfehlen sich hierarchische Verfahren, die eine schnelle und ausreichend gute Lösung für jede mögliche Clusterzahl liefern. Allerdings haben sie den Nachteil, daß einmal getroffene Zusammenfassungen nicht mehr rückgängig gemacht werden. Deshalb bestimmt man zwar aufgrund des hierarchischen Verfahrens die Clusterzahl, verwendet dann aber das gefundene Ergebnis als Anfangszuordnung eines partitionierenden Verfahrens, das (in der Regel) unglückliche Zuordnungen wieder beseitigt.

3.2.5. Parameterversorgung

Die wichtigsten Parameter, die im Sinne einer Sensitivitätsanalyse variiert werden können, sind die Gewichte der einzelnen Merkmale und die Zahl der zu bildenden Cluster.

Als Standardoption (default option) ist in ANCA Gleichgewichtung eingestellt. Wie in Abschnitt 3.2.3. beschrieben, können Gewichte zwischen 1 und 9 für jedes Merkmal subjektiv vergeben werden. Zwar existieren Vorschläge für die automatische Errechnung von "objektiven" Gewichten, etwa um die unterschiedlichen Spannweiten der

Merkmale auszugleichen¹⁾; die Forschung liefert hierzu allerdings eher gegensätzliche Auffassungen; auf eine Implementierung wurde deshalb verzichtet. Bei der Vergabe der Gewichte wird der Anwender unterstützt, sobald ein erstes Ergebnis vorliegt. Für jedes Merkmal wird die Trennfähigkeit bestimmt; unbefriedigende Werte weisen auf die Notwendigkeit hin, das entsprechende Gewicht zu erhöhen.

Falls ein einzelnes Merkmal die Klassifikation dominiert, ist zu erwägen, es versuchsweise von einem Durchgang auszuschließen, d. h. ein Gewicht von Null zu vergeben.

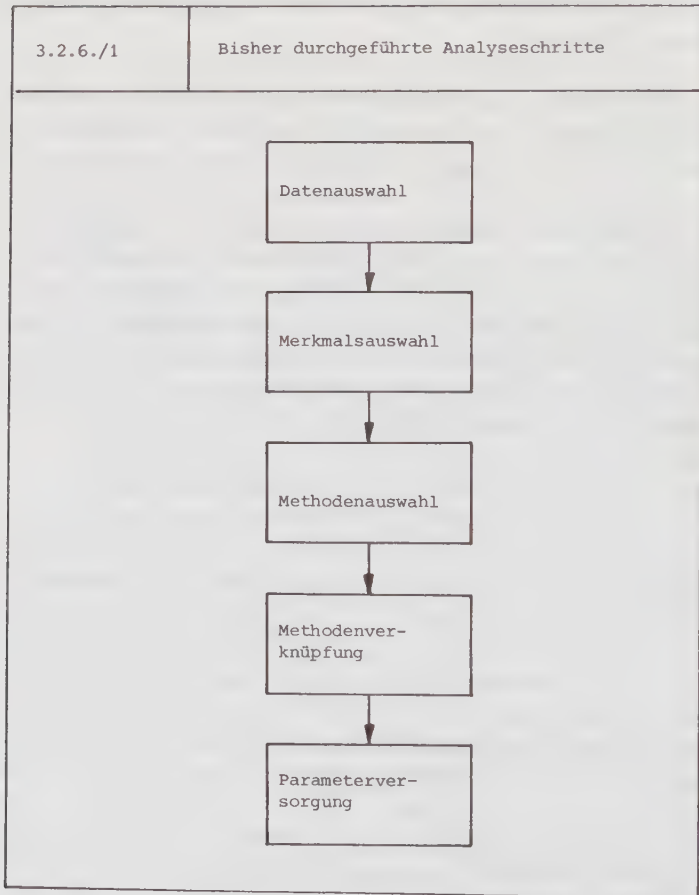
Eines der meist diskutierten Probleme der Clusteranalyse ist die Bestimmung der richtigen Clusterzahl. Der Erfolg der hierarchischen Verfahren mag weitgehend damit zusammenhängen, daß hierfür keine brauchbare Lösung, etwa im Sinne eines "Tests auf Gruppenzahl" existiert. Dementsprechend kann auch in ANCA keine Empfehlung gegeben werden, welche Clusterzahl "richtig" ist. Bei hierarchischen Verfahren wird auf die Zahl hingewiesen, die nach dem jeweiligen Algorithmus am sinnvollsten erscheint; allerdings handelt es sich dabei bereits um eine Interpretationshilfe aufgrund von summarischen Ergebnissen der Clusteranalyse (vgl. Abschnitt 3.2.7.).

Bei partitionierenden Verfahren ist der Anwender allein für die Bestimmung der Clusterzahl verantwortlich; er kann statt einer Zahl einen Bereich angeben, für den Ergebnisse errechnet werden sollen.

¹⁾Vgl. z. B. Wendt, F., a.a.O., S. 38.

3.2.6. Methodendurchlauf

Bisher haben wir in unserer Analyse folgende Schritte durchgeführt (vgl. Abbildung 3.2.6./1).



Wir haben als relevante Stichprobe eine Gruppe von Hausfrauen ausgewählt und mit soziodemographischen Merkmalen sowie Einstellungsdaten beschrieben. Dann wurden die einzelnen Methoden zur Clusteranalyse ausgewählt und verknüpft; für die Variablen wurden noch Gewichte vergeben.

Somit sind alle Informationen vorhanden, die für die Analyse benötigt werden. Nun wird interaktiv der Durchlauf angestoßen; im Normalfall wird auch während der Rechnung der Dialogstatus nicht verlassen, vielmehr stehen die Ergebnisdaten sofort zur Analyse bereit. An dieser Stelle kann auch in den Batchmodus umgeschaltet werden; diese Variante empfiehlt sich vor allem dann, wenn bei vergleichender Analyse die Aufbereitung der Ergebnisse fest vorgegeben ist. Wegen der Laufzeiten ist es nicht notwendig, umzuschalten; die Wartezeiten bleiben im vertretbaren Rahmen.

3.2.7. Interpretation der Ergebnisse

3.2.7.1. Allgemeines

Gerade bei komplizierten Verfahren gilt es, die Ergebnisse angemessen zu präsentieren, um nicht die Akzeptanz des gesamten Verfahrens zu gefährden. Die Erwartungen richten sich dabei im wesentlichen auf folgende Punkte:

- Darstellung des Ergebnisses auf dem Medium und in der Form, die am besten geeignet scheinen
- Herausbilden von Kennzahlen, die das umfangreiche Zahlenmaterial verständlich machen
- Einordnen des Ergebnisses, d. h. Bewertung und Vergleich

Bei der Clusteranalyse muß sich die Präsentation dabei auf

- die gesamte "Clusterlandschaft",
- einzelne Cluster und
- einzelne Merkmale

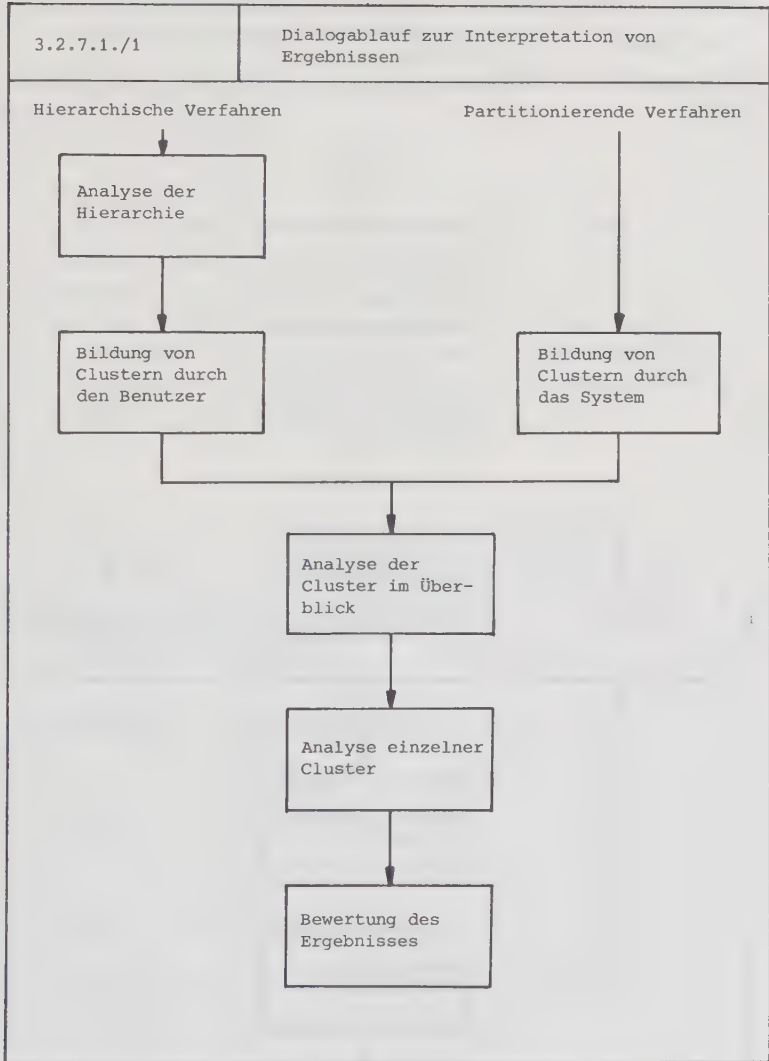
erstrecken. In unserem Beispiel wollen wir prüfen, wie sich Konsumenten auf den Merkmalsraum, nach dem wir sie beurteilen, verteilen; wir wollen uns einige der Gruppen als interessante Zielgruppen für werbliche Aktionen näher ansehen; und wir wollen wissen, auf welche Merkmale die Klassifizierung zurückzuführen ist.

Dementsprechend wurde der in Abbildung 3.2.7./1 dargestellte Ablauf bei der Interpretation geschaffen. Die einzelnen Bausteine werden dabei entsprechend der üblichen Interpretationsfolge abgearbeitet; Variationen des Ablaufs sind bei Bedarf jederzeit möglich.

Die Ergebnisse verwenden benutzereigene Bezeichnungen für Merkmale und Ausprägungen, die der Datendefinitionsdatei entnommen werden. Besonders wichtig ist dies für künstlich in Klassen zusammengefaßte Daten (Altersstufen statt Originalalter) und für nominale Daten (Warengruppe, Einkaufsstätte). Der Anwender kann auch für die gefundenen Cluster eigene Bezeichnungen verwenden, die für weitere Analysen Gültigkeit haben.

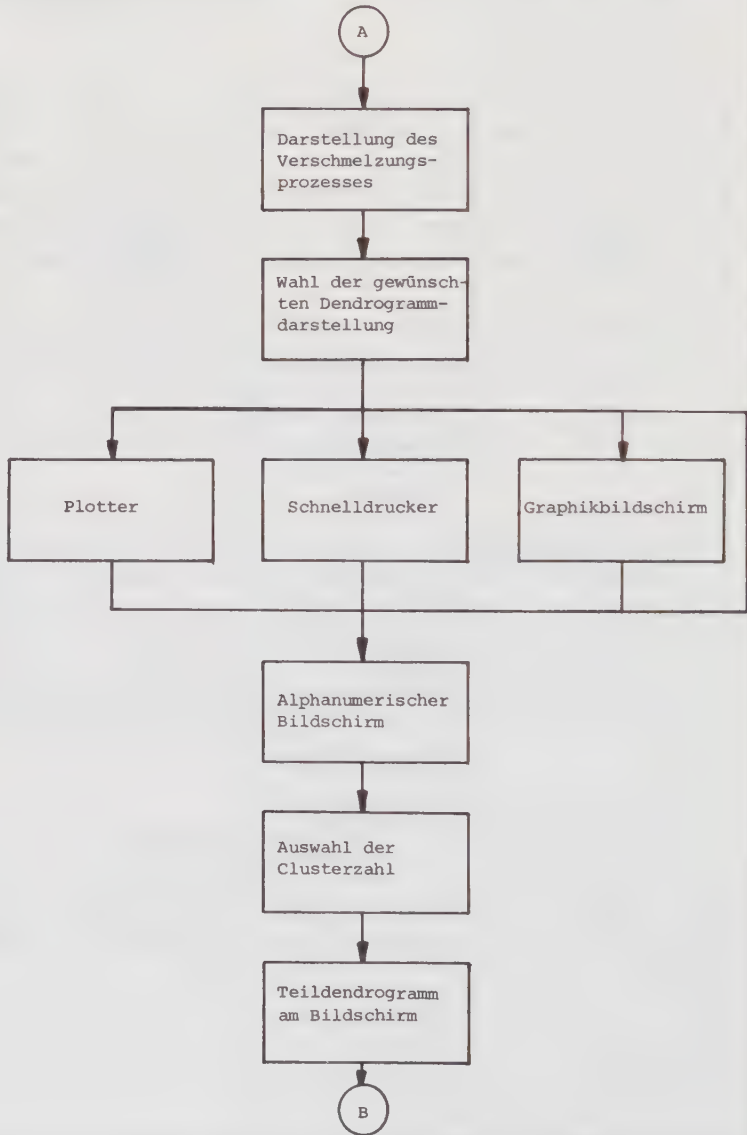
3.2.7.2. Analyse der Hierarchie

Zum Verständnis der Hierarchie sind zwei Hilfen möglich (vgl. Abbildung 3.2.7.2./1):



3.2.7.2./1

Dialogablauf zur Interpretation der
gebildeten Hierarchie



- das Dendrogramm als Darstellungsform für die Hierarchie und
- der Verschmelzungsprozeß, der zeigt, wie sie zustande gekommen ist

Die Darstellung des Verschmelzungsprozesses erfolgt in Anlehnung an Späth; allerdings wird statt der fortlaufenden Zählung der einzelnen Stufen der umgekehrte Weg gegangen, d. h. es wird mitgeteilt, wieviele Cluster auf jeder Stufe gebildet werden (vgl. Abbildung 3.2.7.2./2). Interessant ist hier die Entwicklung des Abstandes, die Aufschluß über eine sinnvolle Zahl von Clustern geben kann.

3.2.7.2./2

Verschmelzungsprozeß bei hierarchischen Clusteranalysen

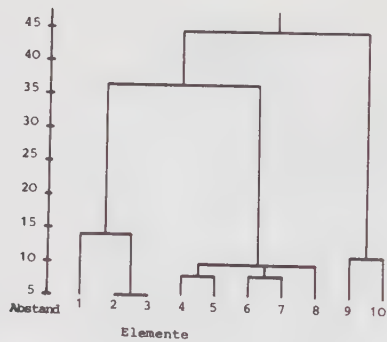
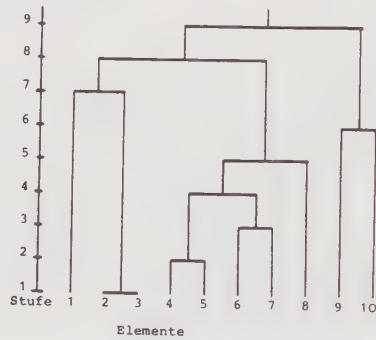
HIERARCHISCHES VERFAHREN NR. 7

CLUSTERZAHL	ABSTAND	EL1	EL2	CLUSTERZAHL	ABSTAND	EL1	EL2
33	0,215	25	29	17	0,433	3	5
32	0,215	12	22	16	0,443	1	7
31	0,218	21	32	15	0,477	12	21
30	0,223	33	42	14	0,488	10	16
29	0,235	8	44	13	0,492	2	24
28	0,243	1	60	12	0,604	33	55
27	0,251	13	14	11	0,618	3	20
26	0,251	24	26	10	0,648	12	53
25	0,256	25	56	9	0,801	2	33
24	0,275	8	11	8	0,805	8	12
23	0,289	2	51	7	0,850	1	4
22	0,290	10	35	6	0,878	3	13
21	0,330	23	40	5	1,101	8	23
20	0,353	4	6	4	1,989	1	3
19	0,358	21	28	3	2,708	8	10
18	0,397	16	25	2	4,586	1	2
SEND.	(HINWEISE DURCH	*INTERPRET')		1	5,492	1	8
\>							

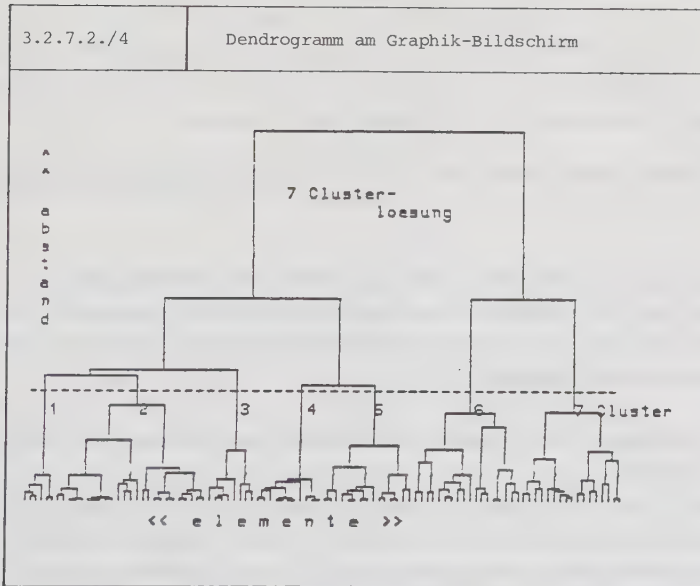
Das Dendrogramm kann wahlweise auf dem Plotter (mit Skalierung nach Abstandssumme oder nach Zuordnungsstufen, vgl. Abbildung 3.2.7.2./3), dem Schnelldrucker oder dem

3.2.7.2./3

Dendrogramm mit Stufen- bzw. Abstands-
skalierung



Bildschirm ausgegeben werden. Steht ein vollgraphisches Terminal zur Verfügung, kann die komplette Hierarchie im Dialog betrachtet und mit den üblichen Graphik-Funktionen (vgl. Abschnitt 4.1.3.) bearbeitet werden. Im Dendrogramm der Abbildung 3.2.7.2./4 wurde beispielsweise durch Beschriftung und Einzeichnen der gestrichelten Linie eine Lösung mit sieben Clustern verdeutlicht.



In jedem Fall ist die Darstellung der (wahlweise) zwei bis zehn letzten Zuordnungsstufen¹⁾ abrufbereit (vgl. Abbildung 3.2.7.2./5). ANCA empfiehlt aufgrund des

¹⁾ In der Darstellung repräsentiert jede Zeile das entsprechende Cluster (Zeile 1 zeigt Cluster 1 usw.) mit den zugehörigen Elementen; im linken Bildteil wird noch dargestellt, wie die Cluster weiter zusammengefaßt würden.

Aufgrund des getrennten Speichers für Graphik und alpha-numerische Ausgabe ist es möglich, die Information der Abbildungen 3.2.7.2./4 und 3.2.7.2./5 übereinanderzu-blenden und gemeinsam darzustellen; neben der generel-len Übersicht ist somit auch noch eine Inhaltsangabe der Cluster in einem Bild gegeben.

3.2.7.3. Ergebnisdarstellung für die Cluster

Nach Wahl der Clusterzahl im Anschluß an ein hierarchi-sches Verfahren, vor allem aber nach Durchführung des partitionierenden Verfahrens oder der Zuordnungsmetho-de, erhält der Benutzer die Möglichkeit, jedes seiner Cluster einzeln oder im Vergleich zu betrachten (vgl. Abbildung 3.2.7.3./1).

Zunächst bietet ein Dendrogramm aller Clusterschwerpunk-te¹⁾ einen allgemeinen Überblick (vgl. dazu Abbildung 3.2.7.3./2). Die Form des Diagramms wurde gewählt, um (maximal) zehn Cluster und sieben Merkmale, d. h. 70 Informationen, zusammen mit Merkmalsbezeichnungen und erklärendem Text auf einer Bildschirmseite darstellen zu können. Die Merkmale sind als Spalten dargestellt²⁾;

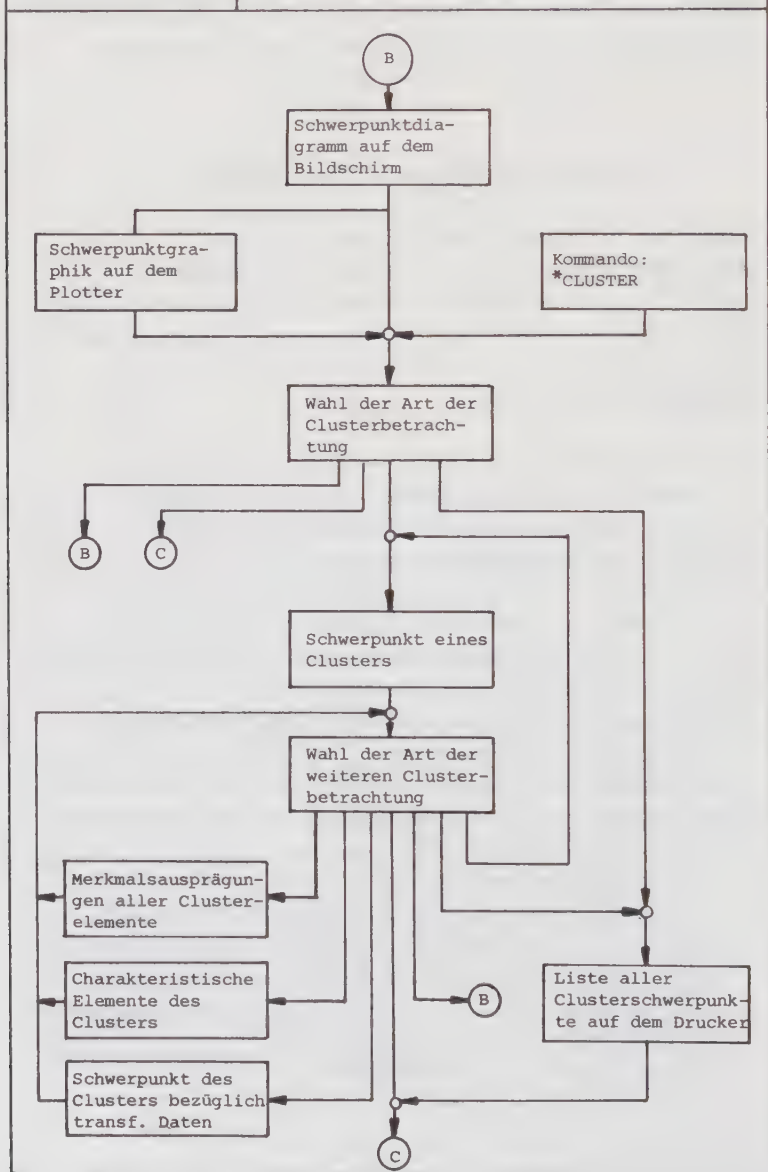
1) Der Begriff Schwerpunkt wurde aus Anschaulichkeits-gründen verwendet; bei nominalen Merkmalen ist dar-unter die häufigste Ausprägung zu verstehen, bei bi-nären der Anteil der Ausprägung 1 an der Gesamtmenge.

2) Die Darstellung unterscheidet sich entsprechend den verschiedenen Skalierungen:

- Bei metrischen Merkmalen wird die Ausprägungsspann-weite in zehn gleich große Intervalle eingeteilt, denen jeweils eine Zeile des Diagramms zugeordnet ist.
- Bei ordinalen und nominalen Merkmalen entspricht je-de Diagrammstufe einer Ausprägung. Um die Gesamtspann-weite deutlich zu machen, wird bei n Ausprägungsstu-fen in der (n + 1)-ten Zeile "-max-" eingetragen.
- Bei binären Merkmalen wird der Anteil der Ausprägung 1 dargestellt; das Intervall zwischen 0 (1 tritt keimnal auf) und 1 (1 tritt ausschließlich auf) wird in zehn gleich große Klassen aufgeteilt.

3.2.7.3./1

Dialogablauf zur Ergebnisdarstellung einzelner Cluster



3.2.7.3./2		Schwerpunktdiagramm der Cluster									

S C H W E R P U N K T E											
= <\											
Z I F F E R N E N T S P R E C H E N C L U S T E R N U M M E R N											
\ =											
MERKMALE	1.0\							12	6		\ 10
1,LEBENSGEMEINSCHAFT	0.9\							680	158		\ 9
2,ALTER	0.8\						1	*	47*		\ 8
3,EINKOMMEN	0.7\	-MAX-	-MAX-	-MAX-				5			\ 7
4,ORTSGROSSE	0.6\	0	1				28	47	30		\ 6
5,LEGE WERT AUF KOSMETIK	0.5\	-MAX-	9	3458	34	56*					\ 5
6,ACHTE AUF MEINE GESUNDEHEIT	0.4\	456	378*	67*	12	470	3				\ 4
7,BIN GEFUEHLSBETONT	0.3\	3789*	45	0	7*						\ 3
	0.2\	0	126	29	680	3	9	2			\ 2
0 = 10.CLUSTER	0.1\	12			59	9		9			\ 1

+ = MEHR ALS 5 GLEICHE											
CLUSTERSCHWERPUNKTE		1	2	3	4	5	6	7			
* = SCHWERPUNKT DER GESAMTMENGE					MERKMALE						
- DURCH EINGABE VON 'P' ERHALTEN SIE EINE GENAUERE PLOTTERZEICHNUNG.											
- DURCH EINGABE VON '*INTERPRET' ERHALTEN SIE INTERPRETATIONSHILFEN.											
EINGABE(P / LEER):											
\>											

die Nummer des Clusters wird in die dem Schwerpunkt entsprechende Position eingetragen¹⁾.

In der Regel ist das Diagramm einfacher lesbar als es hier aufgrund der Einbeziehung des "worst case" erscheint, beispielsweise bei Verarbeitung ausschliesslich binärer Daten. Dennoch wurden Erklärungsbilder geschaffen, die mit einem Kommando abrufbar sind (vgl. Abschnitt 3.2.8.).

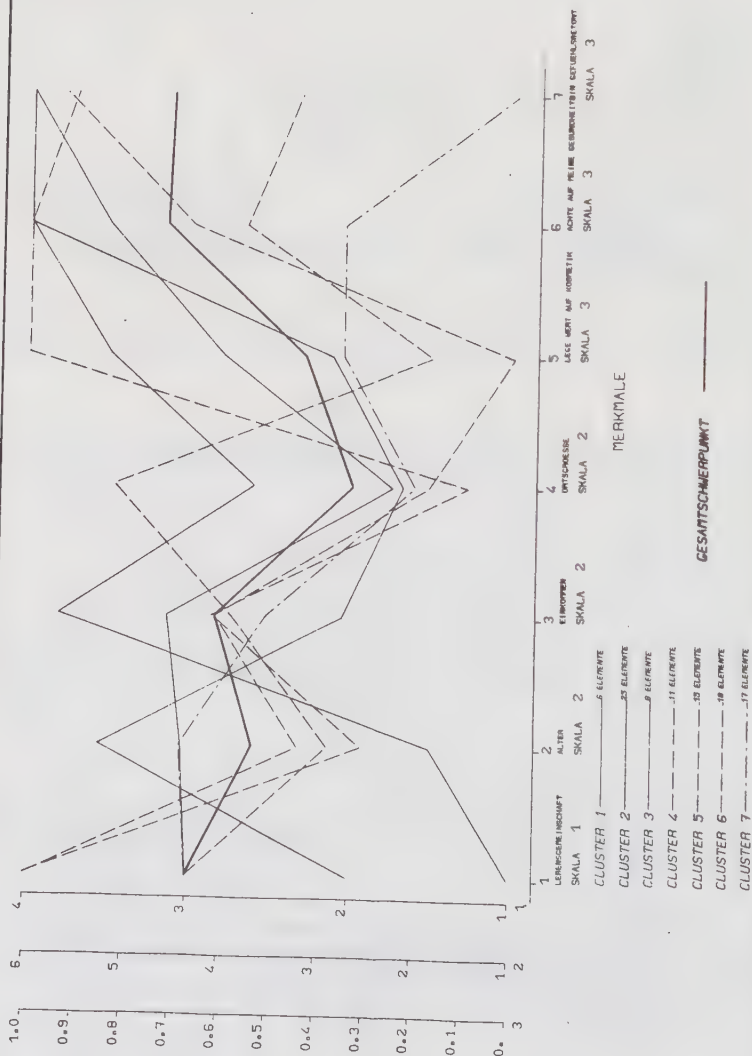
Bessere Möglichkeiten der Darstellung bietet die Ausgabe auf dem Plotter (vgl. Abbildung 3.2.7.3./3). Der Schwerpunkt kann stufenlos eingetragen werden; für je-

1) Sonderfälle:

- 0 kennzeichnet das zehnte Cluster
- + zeigt, daß mehr als fünf Cluster ihren Schwerpunkt bei einer Ausprägung haben
- * kennzeichnet die Lage des Schwerpunktes der Gesamtmenge

3.2.7.3./3

Cluster-Schwerpunkte am Plotter



des Merkmal findet eine eigene Skala Verwendung. Die Gesamtspannweite der Ausprägungsstufen kann somit in der Graphik voll sichtbar gemacht werden. Durch das Verbinden der Punkte über die Merkmale hinweg tritt das "Profil" eines Clusters deutlicher hervor, zumal jede Profillinie aus einer anderen Kombination von Strichart (——, ———, -.-.-,) und Farbe (schwarz, rot, grün) besteht. Bei Einsatz eines graphischen Displays ist diese Darstellung wieder direkt verfügbar, allerdings unter Verzicht auf Mehrfarbigkeit.

Bei unseren Hausfrauen fällt besonders die Randlage des Clusters 1 auf: Es hat die niedrigsten Werte bei "Lebensgemeinschaft" und "Alter" sowie die höchsten bei "Einkommen", "Lege Wert auf Kosmetik" und "Achte auf meine Gesundheit"; offenbar handelt es sich um junge dynamisch-emanzipierte Frauen, die nur eine kleine Gruppe bilden. Die vorhin erkannte Sonderstellung des Clusters 10 ergibt sich auf den ersten Blick vor allem aus dem hohen Alter, während die übrigen Werte eher durchschnittlich sind.

Detaillierter beschrieben werden die Cluster in Einzelbildern. So erkennen wir (Abbildung 3.2.7.3./4), daß alle Mitglieder der Gruppe 1 alleinstehend und ungefähr 20 Jahre alt sind, während der Gesamtaltersdurchschnitt bei 40 Jahren liegt (entsprechend dem ordinalen Wert von 3,66) und nur 12 % der Gesamtmenge alleinstehend sind.

3.2.7.3./4		Schwerpunkt Cluster 1			

CLUSTERNUMMER 1		NACHBARCLUSTER 2			
ANZAHL DER ELEMENTE 7					
MERKMALE		AUSPRAEGUNG	GESAMT	CLUSTER	NACHBAR CLUSTER
1,LEBENS-GEMEINSCHAFT		1=ALLEIN <40J.	12,00 %	100,00 %	100,00 %
2,ALTER		2=20-29 JAHRE	3,66	1,85	1,59
3,EINKOMMEN		6=UEBER 2.STSD	4,06	5,71	2,20
4,ORTSGROESSE		4=MITTELSTADT	2,65	3,85	4,40
5,LEGE WERT AUF KOSMETIK		1=JA	43,00 %	71,42 %	60,00 %
6,ACHTE AUF MEINE GESUNDHEIT		1=JA	72,00 %	100,00 %	100,00 %
7,BIN GEFUEHLSBETONT		1=JA	71,00 %	85,71 %	20,00 %
- EINGABEMOEGLICHKEITEN:		'K'	WEITERE KENNDATEN DIESER CLUSTERS		
		'A'	AUFLISTUNG ALLER ELEMENTE DES CLUSTERS		
		'T'	SCHWERPUNKT BZGL. TRANSFORM. DATEN		
		'L'	AUFLISTUNG ALLER CLUSTER AUF D-DRUCKER		
		'U'	UEBERSICHTSGRAPHIK CLUSTERSCHWERPUNKTE		
		CLUSTERNUMMER	SCHWERPUNKT EINES WEITEREN CLUSTERS		
		LEER	KEIN WEITERES CLUSTER		
EINGABE (K / A / T / L / U / CLUSTERNUMMER / LEER):					
>K					

Cluster werden in der Regel mit zwei verschiedenen Elementen charakterisiert: mit dem Schwerpunkt oder mit demjenigen Objekt, das ihm am nächsten liegt. Beide Elemente sowie das Objekt mit dem größten Abstand zum Zentrum ("Ausreißer") werden mit den Originalausprägungen in einem weiteren Bild dargestellt (vgl. Abbildung 3.2.7.3./5).

Neben der Beschreibung der Cluster sind die enthaltenen Elemente wichtig. ANCA stellt eine Liste zur Verfügung, in der sie nach dem Abstand vom Schwerpunkt geordnet und mit ihren Merkmalen gekennzeichnet werden (vgl. Abbildung 3.2.7.3./6). Dabei kann ein beschreibendes "passives" Merkmal mit ausgegeben werden; dies kann der Name eines Konsumenten, seine Markenverwendung oder eine andere aufschlußreiche Zusatzinformation sein.

Alle abgerufenen Dialogausgaben werden automatisch protokolliert; zusätzlich kann das gesamte Ergebnis mit einem Kommando auf den Schnelldrucker angefordert

3.2.7.4. Bewertung der Cluster

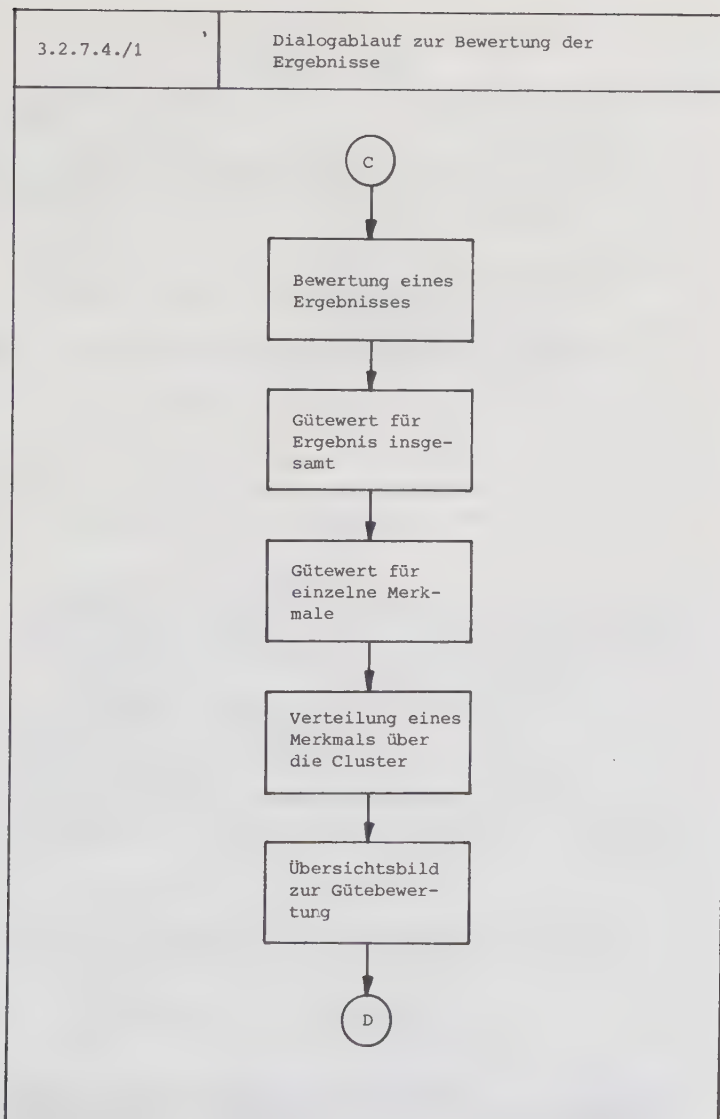
Ein inhaltlich brauchbares Ergebnis muß noch auf seine Güte untersucht werden (vgl. Abbildung 3.2.7.4./1); ein Clusteralgorithmus ohne Kontrolle findet auch in einer nahezu einheitlichen Datenmenge Cluster.

Ein Gütemaß kann nur in Abhängigkeit von der zugrundeliegenden Zielsetzung definiert werden. So ist es für Probleme der Zeichenerkennung wesentlich, daß Buchstaben gefunden oder konzentrische Kreise unterschieden werden können; im wirtschaftswissenschaftlichen Anwendungsbereich wären solche Cluster kaum zu interpretieren.

Das benutzte Gütemaß geht deshalb von varianzanalytischen Grundgedanken aus, kann aber auch auf nicht-metrische Daten angewandt werden. Wesentlich ist, daß eine Einordnung des Ergebnisses in die Kategorien "gut" oder "schlecht" auf den ersten Blick möglich ist; deshalb wurde das Maß auf den Bereich 0 bis 100 normiert; ein niedriger Wert zeigt dabei homogene Cluster an:

$$\text{F-Wert} = \frac{\text{Summe aller Abstände der Elemente zu ihrem Clusterschwerpunkt}}{\text{Summe aller Abstände der Elemente zum Gesamtschwerpunkt}}$$

Der F-Wert des Ergebnisses der Abbildung 3.2.7.4./2 beträgt 57, das Resultat ist also als durchschnittlich zu bezeichnen.



3.2.7.4./2	F-Werte der Cluster
------------	---------------------

ALLE ZAHLENWERTE, DIE IM FOLGENDEN DIE GUETE IHRES ERGEBNISSES ZEIGEN SOLLN, SIND WEGEN DER VERSCHIEDENHEIT DER SKALIERUNGSARTEN UND -SPANNWEITEN AUF DAS INTERVALL 70 - 100/ ABGEBILDET WORDEN. SIE HABEN ABER - WIE SCHON DIE BEZEICHNUNGEN ANDEUTEN - AENHLICHE AUSSAGEKRAFT WIE DIE ENTSPRECHENDEN WERTE DER VARIANZ- UND DISKRIMINANZANALYSE.

CLUSTERZAHL	7	CLUSTERNUMMER	1	2	3	4	5	6	7
ELEMENTE/GESAMT	100	ELEMENTE/CLUSTER	6	25	8	11	15	18	17
MITTL. ABSTAND	32	MITTL. ABSTAND	10	17	16	7	14	30	24

(NORMIERTER ABSTAND DER ELEMENTE ZUM GESAMT- BZW. CLUSTERSCHWERPUNKT)

F-WERT ('VARIANZ IN DEN CLUSTERN'): 57
(MITTLERER ABSTAND DER ELEMENTE ZU IHREM CLUSTERSCHWERPUNKT IM VERHAELTNIS ZUM MITTLEREN ABSTAND VOM GESAMTSCHWERPUNKT)

DURCH 'INTERPRET' ERHALTEN SIE HINWEISE ZUR BEURTEILUNG DIESER WERTE.

SEND,
I>

Als MaB für die Trennschärfe der Cluster und Merkmale wurde der R-Wert implementiert¹⁾, der in seiner Idee dem Clusteranalyse-Programm SYAT²⁾ entnommen wurde. Er mißt die Heterogenität der einzelnen Merkmale zwischen den Clustern und die durch die Clusterbildung erklärte Varianz. Gegenüber der im Programm SYAT implementierten Fassung wurde der R-Wert in ANCA um die Anwendbarkeit auf nominale Daten erweitert und anders normiert.

Die Formel für den R-Wert bei metrischen und ordinalen Daten lautet:

1) Schaller, Klaus, a.a.O., S. 47 f.

2) Stadtler, K., Zur Methodik des Infratest-Typologie-Programms SYAT, interne Firmenschrift, München 1975, S. 17 ff.

$$R_j = \frac{\sum_{c=1}^C n_c (\bar{x}_{cj} - \bar{x}_{gj})^2}{\sum_{i=1}^N (x_{ij} - \bar{x}_{gj})^2} \cdot 100$$

Formel für den R-Wert bei binären und nominalen Daten:

$$R_j = \frac{1}{A_j} \cdot \sum_{k=1}^{A_j} \frac{\sum_{c=1}^C n_c (p_{kcj} - p_{kgj})^2}{N p_{kcj} (100 - p_{kcj})} \cdot 100$$

mit

R_j = R-Wert des j-ten Merkmals

C = Anzahl der Cluster

N = Anzahl der Elemente

n_c = Anzahl der Elemente im c-ten Cluster

A_j = Anzahl der Ausprägungsstufen des j-ten Merkmals

x_{ij} = Ausprägung des j-ten Merkmals beim i-ten Element

$\bar{x}_{cj}$ = Mittelwert des j-ten Merkmals im c-ten Cluster

$\bar{x}_{gj}$ = Mittelwert des j-ten Merkmals in der Gesamtheit

p_{kcj} = Prozentsatz der k-ten Ausprägungsstufe des j-ten Merkmals im c-ten Cluster

p_{kgj} = Prozentsatz der k-ten Ausprägungsstufe des j-ten Merkmals in der Gesamtheit

Abbildung 3.2.7.4./3 zeigt, daß für jedes Merkmal und für die Lösung insgesamt ein R-Wert errechnet wird; so läßt sich z. B. die relativ starke Trennschärfe des

3.2.7.4./3	R-Werte der Lösung

R - W E R T E : (ANTEIL DER DURCH DIE CLUSTERBILDUNG ERKLAERTEN VARIANZ)	
-MASS FÜR DIE HOMOGENITÄT INNERHALB DER CLUSTER BZGL. EINES MERKMALS	
-MASS FÜR DEN BEITRAG DER MERKMALE ZUR CLUSTERBILDUNG	
MERKMALE	R-WERT
1, LERENGEMEINSCHAFT	56 %
2, ALTER	41 %
3, EINKOMMEN	23 %
4, ORTSGROSSE	66 %
5, LEGE WERT AUF KOSMETIK	38 %
6, ACHE AUF MEINE GESUNDHEIT	25 %
7, BIN GEFÜHLSBETONT	69 %
	MITTLERER R-WERT: 46 %
DURCH "INTERPRET" ERHALTEN SIE HINWEISE ZUR BEURTEILUNG DIESER WERTE.	
DURCH EINGABE DER NUMMER EINES MERKMALS ERHALTEN SIE KENNZAHLEN WIE Z.B. DEN	
MITTELWERT DER AUSPRÄGUNGEN DIESES MERKMALS IN DEN EINZELNEN CLUSTERN.	
EINGABE (MERKMALNUMMER / LEET):	
>	

Statements "Bin gefühlsbetont" ebenso zeigen wie die eher geringe der Merkmale "Einkommen" und "Achte auf meine Gesundheit".

Zur Unterstützung dieser Aussagen ist es möglich, die Verteilung der Merkmale in den einzelnen Clustern zu analysieren; beispielsweise wird bei nominalen Daten die exakte Verteilung (Abbildung 3.2.7.4./4) präsentiert.

Eine Übersicht schließt den Modul "Gütebewertung" ab. Dabei werden die kompaktesten Cluster und die Trennschärfe der Merkmale noch einmal herausgehoben. Hinweise für eine veränderte Auswahl der Bausteine zur Verbesserung der Lösung werden aus den Ergebnissen abgeleitet; sie können aber nur grobe Anhaltspunkte geben (Abbildung 3.2.7.4./5).

3.2.7.4./4	Verteilung bei einem nominalen Merkmal
------------	--

MERKMAL: LERNSGEMEINSCHAFT
 SKALIERUNG: NOMINAL

	ANZAHL ELEMENTE	1	2	3	4
GESAMT	100	12 %	13 %	42 %	33 %
CLUSTER 1	6	100 %	0 %	0 %	0 %
CLUSTER 2	25	0 %	0 %	96 %	4 %
CLUSTER 3	8	13 %	88 %	0 %	0 %
CLUSTER 4	11	0 %	0 %	0 %	100 %
CLUSTER 5	15	0 %	0 %	7 %	93 %
CLUSTER 6	18	22 %	11 %	39 %	28 %
CLUSTER 7	17	5 %	24 %	59 %	12 %

BEZEICHNUNGEN DER AUSPRAEGUNGSTUFEN:

1=ALLEIN <40J. 2=ALLEIN >40J. 3=EHE O.KINDER 4=EHE M.KINDER

EINGABE (MERKMALNUMMER / LEER):

>

3.2.7.4./5	Übersicht zur Gütebewertung
------------	-----------------------------

ZUM ABSCHLUSS DER GÜTEBEWERTUNG EINE KURZE ZUSAMMENFASSUNG:

F-WERT ('VARIANZ IN DEN CLUSTERN'): 57
 R-WERT ('ANTEIL ERKLÄRTE VARIANZ'): 46 %
 DIE DREI HOMOGENSTEN CLUSTER: 1 3 4

DAS MERKMAL BIN GEFÜHLSBEDINGT TRUG AM MEISTEN,
 DAS MERKMAL EINKOMMEN AM WENIGSTEN ZUR HOMOGENITÄT DER
CLUSTER BEI.

SIE HABEN NUN FOLGENDE MOGLICHKEITEN:

- DURCH EINGABE EINES 'C' KÖNNEN SIE JEDES DER CLUSTER NOCH EINMAL UNTERSUCHEN;
- DURCH LEERE EINGABE KEHREN SIE ZUR MENÜKARTE DER ERGEBNISBEHANDLUNGEN ZURÜCK

VORSCHLÄGE ZUR VERBESSERUNG DES ERGEBNISSES:

- BEI INHOMOGENEN CLUSTERN ==> ANDERES ABSTANDSMASS
- BEI EINZELNEN INHOMOGENEN CLUSTERN ==> ANDERE CLUSTERANZAHL
- BEI NICHT AUSSAGEKRÄFTIGEN CLUSTERN ==> ANDERES CLUSTERVERFAHREN
- BEI MERKMALEN MIT GERINGEM R-WERT ==> ANDERE GEWICHTUNG DER MERKMALE

EINGABE (C / LEER):

>

3.2.8. Informationshilfen

Informationen zur Unterstützung des Anwenders wurden nach Möglichkeit direkt in den Dialogablauf eingebaut, da passive Hilfen häufig nicht in Anspruch genommen werden¹⁾. Dennoch wird zusätzlich eine HILFE-Funktion²⁾ angeboten für:

- Erklärung von Begriffen, die nicht mehr geläufig sind, im Dialog aber benötigt werden

Beispielsweise kann sich der Anwender noch einmal genau über die Abstandsmaße TANIMOTO und NOMIN informieren, bevor er seine Entscheidung über ein Maß für binäre Daten trifft; der Aufruf erfolgt dabei über den Begriff oder eines seiner Synonyme

- Erklärung von Bildern, die wegen der geringen Bildschirmkapazität sehr kompakt wurden

Beispielsweise existiert ein Erklärungsbild, das die einzelnen Komponenten des Schwerpunktdiagramms verdeutlicht

- Übersicht über die verwendeten Merkmale bzw. Faktoren

Der Anwender kann eine Übersicht über seine Merkmale abrufen; neben Skalenniveau, Zahl der Ausprägungsstufen und Gewichtungsfaktoren sind die einzelnen Ausprägungen aufgeführt. Arbeitet er mit Faktoren statt mit Merkmalen (vgl. Abschnitt 3.4.), kann er ein entsprechendes Bild abrufen, in dem beschrieben ist, welche Merkmale besonders hoch mit den einzelnen Faktoren korreliert sind

¹⁾ Vgl. Möllhoff, L., Unvollkommenes Informationsnachfrage-Verhalten im Mensch-Maschine-Dialog, Stuttgart 1978, S. 149.

²⁾ Vgl. Mertens, P. u.a., Verknüpfung von ..., a.a.O., S. 326 ff.

- Übersicht über die verfügbaren Dialogkommandos

Der Anwender kann diese Hilfe in Anspruch nehmen, wenn er sich "verrannt" hat oder seine Strategie ändern will (vgl. Abschnitt 4.3.2.).

Unmittelbar nach dem HILFE-Ruf wird der letzte Dialogstand wieder hergestellt.

Die Informationshilfen sind kein Ersatz für Methodenwissen; es wäre nicht sinnvoll, Anwender ohne Vorkenntnisse im Dialog mit dem nötigen Know-How für eine erfolgreiche Problemlösung zu versehen.

Insgesamt sind die Erfahrungen mit dieser Funktion positiv. Allerdings ist der erforderliche Speicherbedarf für die Texte nicht zu vernachlässigen; bei beschränktem Speicherplatz ist deshalb der Ersatz durch ein Benutzerhandbuch zu erwägen.

3.3. Systemablauf bei mehreren (vergleichen) Analysen

3.3.1. Anwendungsmöglichkeiten

Vergleiche von Ergebnissen der Clusteranalyse sind vor allem bei folgenden Anwendungsfällen sinnvoll:

- Dynamische Analysen¹⁾

Clusteranalysen, die zu einem Zeitpunkt auf bestimmte Produktmärkte angewandt werden, liefern eine Strukturierung dieses Marktes. Solche Betrachtungen sind in der Regel nur kurzfristig gültig; die Dynamik des Geschehens und die Zufälligkeiten bei der Erhebung erlauben keine exakten Aussagen über eine längere Zeit hinweg. Führt man solche Analysen periodisch durch, können Einsichten in Konstanz bzw. Wanderung gewonnen werden. Ursachen für Wanderungsbewegungen können beispielsweise der Eintritt neuer Produkte, Veränderungen im äußeren Erscheinungsbild bei Artikeln (Relaunch) oder - bei Einstellungsdaten - Veränderungen im Zielsystem der Konsumenten sein.

Ähnliche Untersuchungen können auch für Konsumenten, Unternehmen oder ähnliche über die Zeit relativ konstante Grundgesamtheiten vorgenommen werden.

- Sensitivitätsstudien

Auch bei der Clusteranalyse ist es sinnvoll, bei Unsicherheit von Daten auf eine "exakte" Lösung zu verzichten und dafür die Streubreite des Ergebnisses deutlich zu machen; beispielsweise besteht bei Un-

¹⁾ Vgl. dazu auch Schobert, R., Die Dynamisierung komplexer Marktmodelle mit Hilfe von Verfahren der mehrdimensionalen Skalierung, Dissertation, Nürnberg 1978. Schobert verwendet überwiegend Verfahren der MDS.

sicherheit über die Relevanz der einzelnen Merkmale in einer Analyse das Risiko, daß durch Kompensationswirkungen eine sinnvolle Aussage des Ergebnisses erschwert wird; der bessere Weg ist vielfach, mehrere Analysen mit wechselnden Merkmalen durchzuführen und die Ergebnisse auf identische Mengen ("harte Kerne" der Lösung) zu prüfen. Ebenso sinnvoll ist die Analyse mit verschiedenen Teilmengen oder Verfahren, wenn keine exakten Auswahlkriterien vorliegen¹⁾.

- Interpersonelle Vergleiche

Ergebnisse der Clusteranalyse sind von subjektiven Angaben der Anwender abhängig. Bei gleichen Daten, aber verschiedenen Analytikern ist - wie bei der Nutzwertanalyse - mit verschiedenen Ergebnissen zu rechnen. Es mag nun wiederum sinnvoll sein, aus solchen unabhängig gewonnenen Resultaten ein einziges Ergebnis abzuleiten.

- Vergleich mit einer Idealvorstellung

Häufig haben Planer eine bestimmte Vorstellung - beispielsweise von ihrem Markt -, die mit objektiven Methoden bestätigt werden soll. Dieses Ideal kann mit einer gewollten Marktsegmentierung zusammenhängen, von objektiven Kriterien bestimmt sein oder mit einer Standardtypologie übereinstimmen.

Der Vergleich von Ergebnissen wird von Standardpaketen zur Clusteranalyse nicht unterstützt, obwohl es sich dabei - vor allem bei größeren Datenmengen - um eine recht mühselige Prozedur handelt. Für ANCA wurden deshalb Vergleichshilfen implementiert.

¹⁾ Vgl. dazu auch Schläger, W., Die Klassifikation von Verweildauerhäufigkeiten stationärer Patienten der Medizinischen Klinik der Universität Erlangen-Nürnberg mittels Cluster-Analyse, in: Späth, H. (Hrsg.), Fallstudien ..., a.a.O., S. 81 ff., insbes. S. 92.

3.3.2. Varianten des Ergebnisvergleichs in ANCA

Aufgrund eines Kommandos wird von einem Ergebnis u.a. gespeichert:

- Schwerpunkte der einzelnen Cluster und der Gesamtmenge
- mittlere Abweichung von den Schwerpunkten
- Aufteilung der Elemente auf die einzelnen Cluster

Diese Informationen lassen sich je nach Datenkonstellation in unterschiedlichem Maße zum Ergebnisvergleich nutzen:

- Merkmale und Objekte sind identisch

Dieser Fall liegt beispielsweise bei Analyse von Panneldaten, eingeschränkt bei Marktbetrachtungen sowie bei Vergleichen mit einer Ideallösung vor. Es kann ein Vergleich der Schwerpunktbildung und eine Analyse durchgeführt werden, die die Zugehörigkeit von Objekten zum gleichen Cluster in verschiedenen Ergebnissen untersucht.

- Objekte sind identisch, Merkmale verschieden

Dieser Fall liegt vor, wenn Sensitivitätsstudien mit wechselnden Merkmalen durchgeführt werden. Ein Vergleich ist nur auf der Ebene der Objekte sinnvoll, ein Schwerpunktvergleich ist nicht möglich.

- Objekte sind verschieden, Merkmale sind identisch

Dieser Fall liegt vor, wenn verschiedene Stichproben miteinander verglichen werden sollen oder wenn die Objektmenge, die in eine dynamische Betrachtung eingeht,

starken Veränderungen unterliegt (beispielsweise der Markt für Kleinrechner in den letzten Jahren). Hier ist es nur möglich, einen Schwerpunktvergleich durchzuführen; ein Vergleich der Objektmengen kann bei kleineren Unterschieden mit Hilfe von unterstützenden Maßnahmen erfolgen.

- Objekte und Merkmale sind verschieden

Ein automatischer Vergleich ist nicht sinnvoll.

Für die verschiedenen Konstellationen ist es erforderlich, den jeweils vorhandenen Informationsgehalt soweit als möglich auszunutzen; es müssen deshalb auch verschiedene Vergleichsmethoden geschaffen werden.

Bei allen Ergebnissen, die aus den gleichen Objekten entstanden sind, werden in einer Matrix die Übereinstimmungen von Elementen pro Cluster eingetragen. Das Beispiel der Abbildung 3.3.2./1 zeigt, daß die vier Elemente des Clusters 1 von Ergebnis 1 auch Cluster 1 des Ergebnisses 2 bilden, während Cluster 2 aufgeteilt wird auf die Cluster 2 und 3 des Ergebnisses 2.

3.3.2./1		Übereinstimmung von Elementen in den Clustern verschiedener Ergebnisse			
Ergebnis 1	Elemente in den Clustern		Ergebnis 2		
			(4)	(2)	(4)
	Cluster-nummer		1	2	3
	(4)	1	4	0	0
	(6)	2	0	2	4

Je nach Art der Ergebnisse sind verschiedene Erkenntnisse möglich. Ist die Matrix gleichmäßig besetzt, sind keine konsistenten Gruppierungen vorhanden; viele Leerstellen in der Matrix deuten dagegen eine weitgehende Übereinstimmung der Ergebnisse an. Interessant sind insbesondere bei einer Wanderungsanalyse Aufteilungen und Zusammenfassungen von Clustern. Durch eine anschließende detaillierte Betrachtung der Clusterschwerpunkte läßt sich untersuchen, welche unterschiedlichen Ausprägungen eines Merkmals diese Veränderungen bewirkt haben.

Darüber hinaus können auch die "ähnlichsten" Cluster ermittelt werden. Dabei wird die relative Anzahl der Elementübereinstimmungen geprüft; zwei Cluster sind dann als sich entsprechendes Paar definiert, wenn eine von folgenden Bedingungen erfüllt ist:

1. Mindestens 80 % der Elemente des einen Clusters sind auch Mitglieder des anderen Clusters. Damit können auch kleine Cluster, die von größeren aufgenommen werden, erfaßt werden. Einzelne Cluster treten dadurch unter Umständen mehrmals auf.
2. Mindestens 60 % der Elemente des einen Clusters und 40 % der Elemente des anderen Clusters stimmen überein. Damit sollen kleinere Unschärfen am Rande der Cluster berücksichtigt werden.

Die o. a. Zahlenwerte sind empirisch nicht abgesichert, ihre Verwendung führte aber in den Tests meistens zu Resultaten, die einer (subjektiven) Einschätzung der Cluster entsprachen.

Nach dem Vergleich von zwei Ergebnissen lassen sich weitere Ergebnisse zuordnen. Auf diese Weise können einzelne konsistente Elementgruppierungen über mehrere Ergebnisse verfolgt werden.

Bei einem Schwerpunktvergleich wird der Abstand zwischen allen Schwerpunkten errechnet; das Minimum dieser Werte bestimmt die größte Ähnlichkeit. In diesem Fall ist es erforderlich, die einzelnen Schwerpunktkoordinaten, die Streuung und die Zahl der Elemente innerhalb der Cluster zusätzlich zur Beurteilung heranzuziehen.

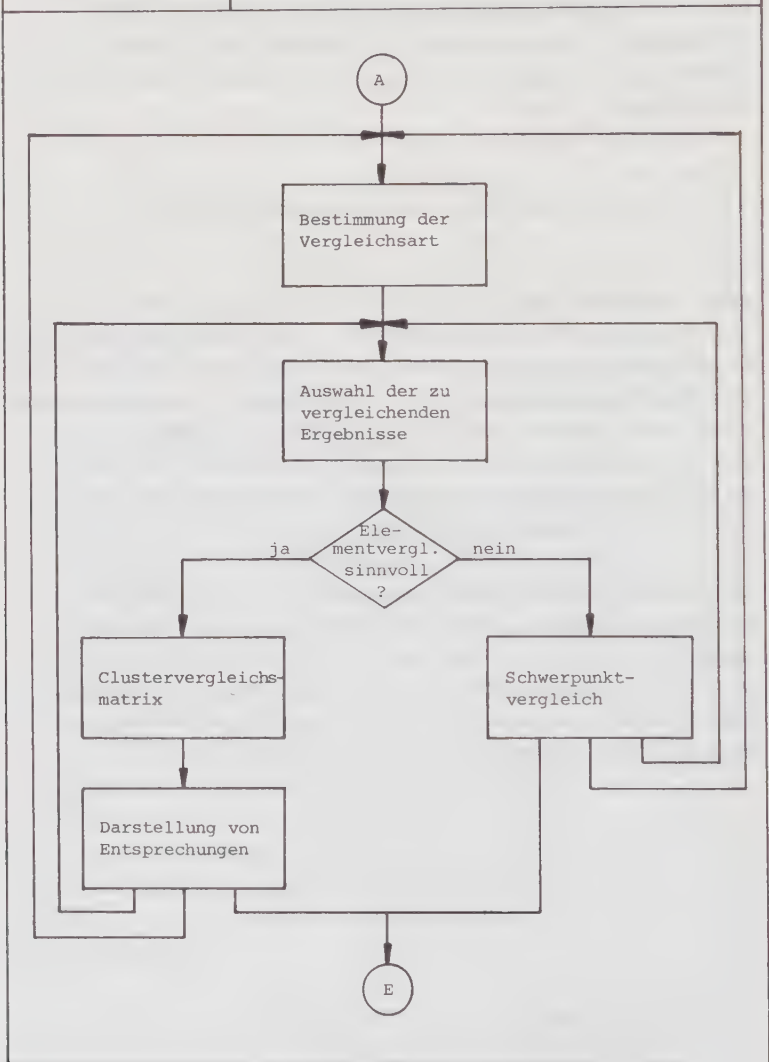
3.3.3. Dialogablauf bei vergleichender Analyse

Der Dialogablauf bei vergleichender Analyse ist linear gestaltet (vgl. Abbildung 3.3.3./1). In der ersten Phase teilt der Anwender dem System mit, welche Ergebnisse gespeichert sind, und fragt nach der gewünschten Vergleichsart (vgl. Abbildung 3.3.3./2). Eine automatische Auswahl ist nicht möglich, der Anwender muß seine Strategie selbst festlegen. In unserem Fall wurde mit fünf unterschiedlichen Anfangszuordnungen gearbeitet; wir wählen Variante 1.

Im nächsten Schritt beschreibt ANCA die vorliegenden Ergebnisse genauer; neben Informationen über Zahl der Merkmale und Cluster und über die Güte der Ergebnisse stellt das System auch grobe Aussagen über die Cluster bereit (vgl. Abbildung 3.3.3./3). Der Anwender entscheidet sich für zwei Resultate, im Beispiel für 3 und 4.

3.3.3./1

Dialogablauf bei vergleichender Analyse



3.3.3./2	Darstellung der möglichen Vergleichsarten
----------	---

FOLGENDE GESICHERTEN ERGEBNISSE STEHEN IHNEN ZUR VERFUEGUNG:

- 1 ANFZUORD1
- 2 ANFZUORD2
- 3 ANFZUORD3
- 4 ANFZUORD4
- 5 ANFZUORD5

SIE HABEN FOLGENDE MOEGLICHKEITEN DES VERGLEICHS VON ERGEBNISSEN:

- (1) VERSCHIEDENE DATENAUFBEREITUNGEN, ABSTANDSDEFINITIONEN, CLUSTERUNGSVERFAHREN, CLUSTERANZAHLN ==> 'BESTES' ERGEBNIS, 'ECHTE' CLUSTER
- (2) VERSCHIEDENE MERKMALE ==> KONSISTENTE GRUPPIERUNGEN VON ELEMENTEN
- (3) VERSCHIEDENE STICHPROBEN AUS DER ELEMENTMENGE ==> SENSITIVITAET DES ERGEBNISSES
- (4) VERSCHIEDENE DATENERHEBUNGEN DERSELBEN ELEMENTE ==> WANDERUNGSBEWEGUNGEN
- (5) HINWEISE ZU DIESEN ALTERNATIVEN

EINGABE:
I>1

3.3.3./3	Darstellung der vorliegenden Ergebnisse
----------	---

BEZEICHNUNG	1-ANFZUORD1	2-ANFZUORD2	3-ANFZUORD3	4-ANFZUORD4	5-ANFZUORD5
ANZAHL CLUSTER	7	7	7	7	5
ANZAHL MERKMALE	7	7	5	5	5
F-WERT	42	42	29	31	44

ANZAHL ELEMENTE / MITTL. ABSTAND ZUM GESAMT- BZW. CLUSTERSCHWERPUNKT

GESAMT	100 / 20	110 / 20	100 / 20	100 / 20	110 / 20
CLUSTER 1	14 / 8	16 / 8	11 / 5	24 / 6	25 / 6
CLUSTER 2	19 / 7	23 / 7	12 / 5	10 / 4	37 / 10
CLUSTER 3	16 / 10	17 / 9	17 / 3	11 / 5	12 / 5
CLUSTER 4	15 / 9	16 / 9	13 / 3	13 / 3	26 / 12
CLUSTER 5	15 / 6	16 / 6	13 / 12	22 / 5	10 / 7
CLUSTER 6	7 / 9	7 / 9	10 / 4	9 / 7	
CLUSTER 7	14 / 10	15 / 10	24 / 6	11 / 13	
CLUSTER 8					
CLUSTER 9					
CLUSTER 10					

= DURCH EINGABE ZWEIER ERGEBNISNUMMERN KOENNEN SIE DIESE ERGEBNISSE VERGLEICHEN:
= DURCH LEERE EINGABE VERLASSEN SIE DEN ERGEBNISVERGLEICH.
EINGABE (ERGEBNISNUMMERN, GETRENNT DURCH '/' / LEER):
I>3/4

Ist der Elementvergleich zulässig, stellt ANCA die entsprechende Matrix zur Verfügung (vgl. Abbildung 3.3.3./4). In unserem Fall liegt eine sehr gute Übereinstimmung der Ergebnisse vor; Wanderungsbewegungen haben nur zwischen den Clustern 2 und 5 (Ergebnis 3) sowie 5 und 7 (Ergebnis 4) stattgefunden, während vier Gruppierungen völlig identisch sind. Das folgende Bild (Abbildung 3.3.3./5) bestätigt diesen Befund.

3.3.3./4	Vergleichsmatrix für Ergebnis 3 und 4
<pre> ***** GLEICHE ELEMENTE IN DEN CLUSTERN DER ERGEBNISSE ANFZUORD3 UND ANFZUORD4 . LESEBEISPIEL: DAS CLUSTER 3 DES ERGEBNISSES 3 HAT MIT DEM CLUSTER 5 DES ERGEBNISSES 4 17 GEMEINSAME ELEMENTE. 4-ANFZUORD4 (24) (10) (11) (13) (22) (9) (11) ->-----> 1 2 3 4 5 6 7 3-ANFZUORD3 1-> (11) 1: (12) 2: (17) 3: IN (13) 4: KLAMMERN (13) 5: DIE (10) 6: ELEMENTE (24) 7: 24 / CLUSTER SEND. 1> ***** </pre>	

3.3.3./5	Darstellung sich entsprechender Cluster
----------	---

```

*****
FOLGENDE CLUSTER ENTSPRECHEN SICH AM MEISTEN (MEHRFACHES AUFTRETEN = TEILUNG):
3-ANFZUORD3      4-ANFZUORD4
1      -      3
2      -      7
3      -      5
4      -      4
5      -      6
6      -      2
7      -      1

- DURCH EINGABE EINER ERGEBNISNUMMER KÖNNEN SIE EIN WEITERES ERGEBNIS ZUORDNEN
1-ANFZUORD1      2-ANFZUORD2      3-ANFZUORD3      4-ANFZUORD4      5-ANFZUORD5
- DURCH EINGABE EINER ERGEBNISNUMMER UND - GETRENNT DURCH '/' - EINER
  CLUSTERNUMMER ERHALTEN SIE DIE AUFLISTUNG DES CLUSTERSCHWERPUNKTS.
- DURCH EINGABE EINES 'V' KÖNNEN SIE EINEN WEITEREN VERGLEICH DURCHFÜHREN.
- DURCH LEERE EINGABE VERLASSEN SIE DEN VERGLEICH VON ERGEBNISSEN.
EINGABE (ERGEBNISNUMMER / ERGEBNISNUMMER UND CLUSTERZAHL / V / LEER):
|>
*****

```

An dieser Stelle können wir die Entsprechungen mit einem weiteren Ergebnis anfordern oder - zur Analyse von Abweichungen - Informationen über einzelne Cluster abrufen.

3.4. Systemablauf bei Methodenverbund in MADAS

Mit dem System MADAS wird in der ersten Phase (vgl. Abschnitt 2.2.2.) eine Datenmenge definiert, die untersucht werden soll; in unserem Fall sei es eine Stichprobe von 110 Arbeitsplätzen, die durch Anforderungsprofile beschrieben sind¹⁾.

MADAS bietet die vorliegenden (in unserem Fall 15) Merkmale zur Analyse an (vgl. Abbildung 3.4./1).

¹⁾ Vgl. Herrmann, U., Analyse eines Analytischen Arbeitsplatzbewertungs- und Entlohnungssystems, Diplomarbeit, Nürnberg (in Vorbereitung).

3.4.1	Bild zur Merkmalsauswahl in MADAS
-------	-----------------------------------

MOEGLICHE MERKMALE FUER DIE METHODEANWENDUNG

(1)	AUSBILDUNG
(2)	ERFAHRUNG
(3)	GEISTIGE ANFORDERUNG
(4)	KONTAKTE
(5)	GESCHICKLICHKEIT
(6)	MUSKELBELASTUNG
(7)	REISEN
(8)	EIGENVERANTWORTUNG
(9)	FUEHRUNGSVERANTWORTUNG
(10)	FACHAUFSICHT
(11)	BELASTUNG SINNE&NERVEN
(12)	AUTODAHREN

SIE KOENNEN BIS ZU SIEBEN MERKMALE WAEHLN.
ES FOLGEN NOCH WEITERE MERKMALE.

EINGABE

Für diese Merkmale stellt MADAS Häufigkeitsverteilungen bereit (vgl. Abbildung 3.4./2); zur weiteren Analyse wählen wir sieben Merkmale aus.

3.4.2	Häufigkeitsverteilung für die Datenmenge in MADAS
-------	---

MERKMAL : EIGENVERANTWORTUNG			HERRMANN-TES GESAMTHEIT	
			ABS.	%
(1)	BIS	000008		
(2)	BIS	000017	12	10,91
(3)	BIS	000025	11	10,00
(4)	BIS	000033	19	17,27
(5)	BIS	000042	13	11,82
(6)	BIS	000050	16	14,55
(7)	BIS	000058	14	12,73
(8)	BIS	000067	14	12,73
(9)	BIS	000075	7	6,36
(10)	BIS	000083	1	0,91
(11)	BIS	000092	3	2,73
(12)	BIS	000100		

110

Auf die prinzipiell mögliche Zusammenfassung einzelner Ausprägungen wollen wir bei unseren (metrischen) Daten verzichten.

Bei der Methodenauswahl berücksichtigt MADAS die gespeicherte Information über die Skalierung und schlägt vier multivariate Verfahren vor (vgl. Abbildung 3.4./3). Wir wählen die Clusteranalyse; der Übergang zum System ANCA erfolgt automatisch.

3.4./3	Menu der zulässigen multivariaten Verfahren
SIE HABEN SICH FUER DIE MULTIVARIATE STRUKTURANALYSE ENTSCHIEDEN. ALLE MERKMALE STAMMEN AUS EINER STICHPROBE UND SIND METRISCH SKALIERT. ES STEHEN FOLGENDE METHODEN ZUR VERFUEGUNG: (1) MULTIPLE PARTIELLE KORRELATION (2) MULTIPLE KORPELATION (3) CLUSTERANALYSE (4) FAKTORENANALYSE EINGABE:	

Nachdem ANCA fünf Cluster empfohlen und dargestellt hat (vgl. Abbildung 3.4./4), wollen wir uns die beiden "extremen" Cluster 1 und 5 näher betrachten; wir definieren beide als neue Stichproben für MADAS.

3.4./4		Ergebnis der Clusteranalyse für Arbeitsplatzdaten									

S C H W E R P U N K T E D E R C L U S T E R											
= < Z I F F E R N E N T S P R E C H E N C L U S T E R N U M M E R N =											
MERKMALE		1.0	5								10
1, AUSBILDUNG		0.9	4		5				5		9
2, ERFAHRUNG		0.8		5				5			8
3, GEISTIGE ANFORDERUNG		0.7			4						7
4, GESCHICKLICHKEIT		0.6	*	34*		2					6
5, MUSKELBELASTUNG		0.5	23	2	*	1	1	4*	4*		5
6, EIGENVERANTWORTUNG		0.4		1	23		2	23	123		4
7, BELASTUNG SINNESNERVEN		0.3				4*					3
		0.2	1		1	3	4*	1			2
		0.1				5	35				1

+ = MEHR ALS 5 GLEICHE CLUSTERSCHWERPUNKTE											
* = SCHWERPUNKT DER GESAMTMENGE											
1 2 3 4 5 6 7											
M E R K M A L E											
- DURCH EINGABE VON 'P' ERHALTEN SIE EINE GENAUERE PLOTTERZEICHNUNG.											
- DURCH EINGABE VON '*INTERPRET' ERHALTEN SIE INTERPRETATIONSHILFEN.											
EINGABE(P / LEER) :											
>											

Die Cluster können mit dem gesamten Methodenvorrat analysiert werden; im folgenden sei exemplarisch die vergleichende Häufigkeitsdarstellung von Cluster und Gesamtmenge gezeigt (vgl. Abbildungen 3.4./5 und 3.4./6). In bezug auf "geistige Anforderung" zeigt sich völlige Trennschärfe zwischen beiden Clustern; auch bei einigen anderen Merkmalen ist dieses Ergebnis ähnlich deutlich.

Die Bestimmung der Trennschärfe der Merkmale könnte auch die Diskriminanzanalyse übernehmen, die direkt anschließend an die Clusteranalyse aufgerufen werden kann¹⁾.

¹⁾ Vgl. Schubert, V., Verknüpfung von Faktorenanalyse, Clusteranalyse und multipler Diskriminanzanalyse im Rahmen der interaktiven Methodenbank MADAS, Diplomarbeit, Nürnberg 1977.

3.4./5		Vergleichende Häufigkeitsdarstellung des Clusters "Führung" mit der Gesamtmenge				
MERKMAL : GEISTIGE ANFORDERUNG						
		STICHPROBE		FUEHRUNG		
		ABS.	%	ABS.	%	
(1)	BIS	000008				
(2)	BIS	000017		8	7,27	
(3)	BIS	000025		10	9,09	
(4)	BIS	000033		20	18,18	
(5)	BIS	000042		13	16,36	
(6)	BIS	000050		10	9,09	
(7)	BIS	000054		9	8,18	
(8)	BIS	000067	1	4,35	10	9,09
(9)	BIS	000075	1	4,35	2	1,82
(10)	BIS	000083	10	43,48	11	10,00
(11)	BIS	000092	6	26,09	7	6,36
(12)	BIS	000100	5	21,74	5	4,55
		23		110		

3.4./6		Vergleichende Häufigkeitsdarstellung des Clusters "Fließband" mit der Gesamtmenge			
MERKMAL : GEISTIGE ANFORDERUNG					
		STICHPROBE		FLIESSBAND	
		ABS.	%	ABS.	%
(1)	BIS	000008			
(2)	BIS	000017	6	23,08	8 7,27
(3)	BIS	000025	9	34,62	10 9,09
(4)	BIS	000033	9	34,62	20 18,18
(5)	BIS	000042	2	7,69	18 16,36
(6)	BIS	000050			10 9,09
(7)	BIS	000058			9 8,18
(8)	BIS	000067			10 9,09
(9)	BIS	000075			2 1,82
(10)	BIS	000083			11 10,00
(11)	BIS	000092			7 6,36
(12)	BIS	000100			5 4,55
			26	110	

Eine Variante des Ablaufs bringt der Einsatz der Faktorenanalyse vor ANCA¹⁾. Sie untersucht, ob sich die Merkmale (aufgrund von Interkorrelationen) durch (weniger) unabhängige Faktoren erklären lassen, und errechnet für alle Objekte die Faktorwerte. Für ANCA bestehen folgende Möglichkeiten:

- Clusterung mit Faktorwerten, wenn eine sinnvolle Reduktion gefunden wurde
- Clusterung mit den Originalmerkmalen, wenn (weitgehende) Unabhängigkeit festgestellt wurde

Zur Erinnerung daran, wie die Faktoren zusammengesetzt sind, kann an jeder Stelle eine entsprechende Informationshilfe abgerufen werden (vgl. Abbildung 3.4./7); dabei gibt die Anzahl der "+"-oder "-"-Zeichen Auskunft über die Stärke des (positiven oder negativen) Zusammenhangs zwischen einer Variablen und einem Faktor.

3.4./7		Beziehung zwischen Originalmerkmalen und Faktoren		

ZUSAMMENSETZUNG DER FAKTOREN				
NAME DES FAKTORS	MERKMALSNR.	MERKMALNAME	LADUNG	
1 GEIST	0003	GEISTIGE ANFORDERUNG	+++	
	0008	EIGENVERANTWORTUNG	++	
	0011	BELASTUNG SLAVE	+	
2 AUSWAHLER	0013	BELASTUNG MONOTONIE	++	
	0012	AUTOFAHREN	+	
3 KUERPER	0006	MUSKELBELASTUNG	=	
SEND				
\>				

¹⁾ Vgl. Schubert, V., Verknüpfung von ..., a.a.O.

4. Datenverarbeitungstechnische Realisation des Systems ANCA

4.1. Grundkonzept

4.1.1. Bedarfsangaben

ANCA ist auf dem Informatik-Forschungsrechner TR 440 des Regionalen Rechenzentrums Erlangen realisiert worden. Bei der Erstellung wurden die Prinzipien der Modularität, Portabilität und Erweiterbarkeit berücksichtigt.

Die Minimalkonfiguration für ANCA umfaßt eine Wechselplatte, einen Bildschirm oder Fernschreiber und einen Drucker.

ANCA besteht aus einem Operator, der direkt gestartet werden kann. Das Hauptprogramm und die 22 Unterprogramme sind fest montiert; ein dynamisches Zuladen¹⁾ - etwa von Programmen für spezielle Methoden - ist nicht möglich. Die Quellprogramme für ANCA umfassen ca. 4.600 Statements in COBOL und 2.200 Statements in FORTRAN; der Verbund mit MADAS ist dabei nicht berücksichtigt.

Die Verarbeitung von ANCA erfordert

66 k Arbeitsspeicher (KSB)²⁾
81 k Trommelspeicher (TSB)
190 k Plattenspeicher (PSB).

¹⁾ Vgl. Abschnitt 5.3.1.

²⁾ Bei Arbeit mit graphischer Software erhöht sich dieser Bedarf auf 90 KSB.

Der Rechenzeitbedarf beträgt bei einer typischen Analyse 20 Sekunden, bei vergleichenden Untersuchungen oder bei Zuordnung von Elementen erhöht er sich entsprechend.

4.1.2. Aufruf des Systems

Das System steht mehreren Benutzern im Teilhaberbetrieb zur Verfügung. Der Aufruf sollte soweit wie möglich vereinfacht werden, was bei der angesprochenen Zielgruppe besonders wesentlich ist; es ist quasi "auf Knopfdruck" möglich, mit ANCA zu arbeiten.

Der Benutzer muß neben der Anmeldung des Gespräches nur das MADAS-Gedächtnis aufrufen, um eine Analyse durchführen zu können. Innerhalb dieses Gedächtnisses sind zwei Kommandos relevant:

- ein Kommando für den Aufruf von ANCA allein
- ein Kommando für den Aufruf von MADAS, das auch den Verbund mit ANCA ermöglicht; der Übergang geschieht automatisch in beide Richtungen

Der Benutzer muß mit dem jeweiligen Kommando nur seine spezifischen Dateien bekanntgeben; die Dateien, die in jedem Lauf erforderlich sind, werden automatisch bereitgestellt (vgl. Abschnitt 4.2.4.).

4.1.3. Dialog am Graphikbildschirm

Aufgrund der erwähnten Vorteile (Abschnitt 3.2.7.) scheint ein Einsatz des Graphikbildschirms in jedem

Fall sinnvoll. Allerdings kann in einer typischen betrieblichen Konfiguration nicht mit einem solchen Gerät gerechnet werden; um ANCA trotzdem übertragbar zu halten, wurde der Einsatz der Graphik als Wahlmöglichkeit, nicht aber als unbedingte Voraussetzung implementiert. Der Benutzer gibt im ersten Dialogbild bekannt, ob er die Graphikversion beansprucht oder nicht, das System sorgt dann selbst für die notwendigen Aufrufe.

Durch den Aufruf des Graphikpostprozessors im System ANCA können die erstellten Zeichnungen mit geringfügigen Änderungen auf dem Bildschirm erzeugt werden. Um dem Benutzer die Einarbeitung in den Prozessor zu ersparen, wurden die entsprechenden Routinen so geändert, daß allein durch Drücken der "SEND"-Taste die Zeichnungen abrufbar sind. Die Möglichkeiten, die der Prozessor zusätzlich bietet, z. B.

- Ausschnittsvergrößerungen
- Wiederholung und Überspringen von Zeichnungen
- Ausgabe mehrerer Zeichnungen

stehen dem zur Verfügung, der diese Funktionen beherrscht.

Bei dem am Regionalen Rechenzentrum Erlangen verfügbaren Graphiksichtgerät HP 2648A sind für den ANCA-Benutzer folgende Funktionen besonders interessant:

- Hardcopy des Bildschirms
- Einblendung des Alpha-Bildschirms mit dem ANCA-Dialog in die Graphikausgabe und umgekehrt

- "Roll-up" und "Roll-down" des Bildschirminhaltes bild- oder zeilenweise (auch für reine Textausgaben!)
- Off-line-Einfügen von Linien und Text
- Speichern der Graphikausgabe bis zur nächsten Zeichnung (mit beliebiger Abrufmöglichkeit)
- Ausschnittvergrößerung ("Lupe").

4.2. Programmaufbau von ANCA

4.2.1. Sprachkonzept

Bei der Realisierung von ANCA wurde auf Übertragbarkeit großer Wert gelegt. Andererseits mußte für verschiedene Aufgaben das jeweils geeignete Werkzeug gefunden werden. Als Kompromiß ergab sich die Verwendung von COBOL für den Rahmen des Programmsystems und von FORTRAN für rechenintensive Bausteine. Die Einbeziehung von Betriebssystemroutinen oder von rechnerspezifischen Sprachelementen wurde soweit als möglich umgangen, ist aber an manchen Stellen aus Gründen des Benutzerkomforts unvermeidlich.

COBOL wird speziell in jenen Programmen verwendet, die die Aufrufsteuerung der einzelnen Moduln (vgl. Abschnitt 4.2.2.) sowie die Dialogführung beinhalten.

Hier kann der neben der Unabhängigkeit von der Rechenanlage wohl bedeutendste Vorteil von COBOL, das komfortable Texthandling, genutzt werden. Infolge der Vielzahl von Dialogschritten in ANCA ist eine einfache und klare Implementierung der entsprechenden Routinen von großer Bedeutung; deswegen wurde auch der große Programmvolumen, den COBOL mit sich bringt, in Kauf genommen.

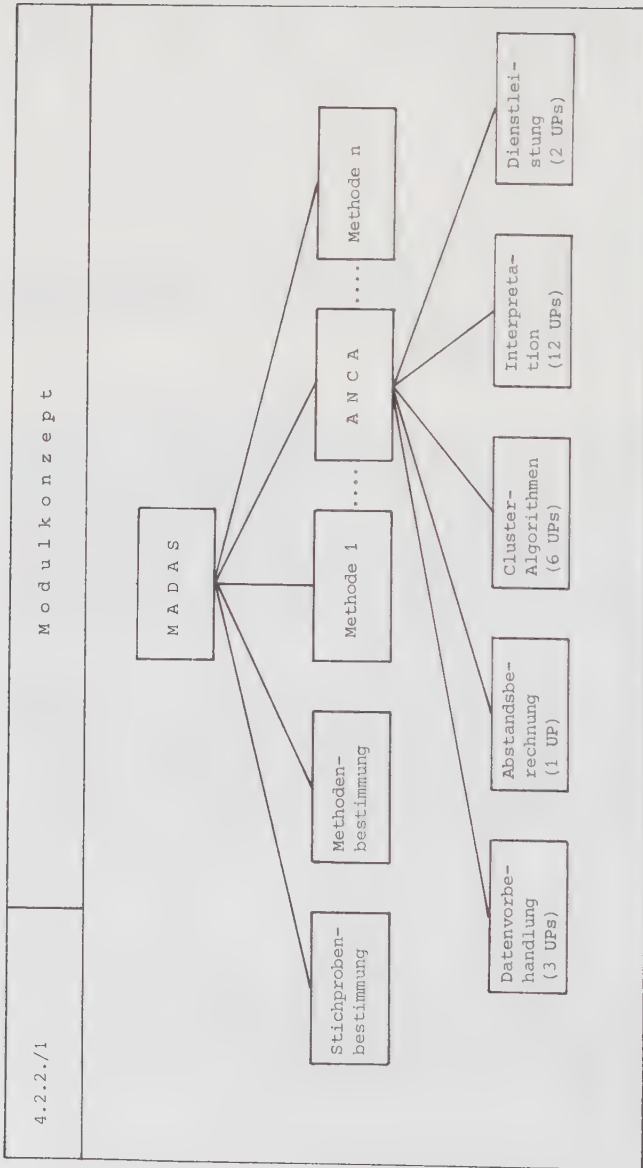
FORTRAN zeichnet sich durch einfachere und schnellere Verarbeitung arithmetischer Operationen aus. In allen Rechenbausteinen von ANCA findet diese Sprache Verwendung. Erforderlich wurde FORTRAN auch für Moduln, die den Plotter ansprechen.

Rechnerspezifische Elemente finden sich vor allem im Zusammenhang mit der Graphiksoftware. Neben den Graphikkommandos des TR 440 wurde auf den Graphik-Postprozessor des RRZE zurückgegriffen. Darüber hinaus wurde die Möglichkeit genutzt, aus dem Programm heraus auf Betriebssystemkommandoebene zu gehen und Kommandos abzusetzen, die Einfluß auf den weiteren Dialogablauf haben.

Im Gegensatz zu interpretativen Systemen ist in ANCA die Implementierungssprache deutlich von der Benutzersprache getrennt; Anwender kommunizieren mit dem System nur über Dialogeingaben oder mit speziellen Kommandos (vgl. Abschnitt 4.3.).

4.2.2. Modulkonzept

Das Modulkonzept von MADAS, das - neben anderen Methodenprogrammen - auch ein selbständiges System zur Clusteranalyse beinhaltet, wird in ANCA hierarchisch weitergeführt (vgl. Abbildung 4.2.2./1).



ANCA besteht aus einem Hauptprogramm, das auch die zentrale Steuerung übernimmt, und 22 Unterprogrammen. Die Abgrenzung der Unterprogramme erfolgt in Anlehnung an die Teilaufgaben der Clusteranalyse; zwei spezielle Dienstleistungsmoduln sind für Dialogausgaben und Prüfung der Eingaben zuständig.

Die Unterprogramme sind auf eine Aufgabenstellung zugeschnitten und übersichtlich gestaltet. Im Hauptprogramm wurde eine weitere Modularisierung durchgeführt: Die Aufrufumgebung der Unterprogramme und die nichtauslagerbaren Teile wurden in einzelne Sections zusammengefaßt und entsprechend gekennzeichnet (z. B. PART SECTION, VERGLEICH SECTION).

Neben Isolierung der Teilaufgaben und Übersichtlichkeit gewährleistet diese Modularisierung Offenheit des Systems gegenüber neuen Algorithmen und sonstigen (z. B. Interpretations-)Bausteinen.

4.2.3. Steuerkonzept

Der Benutzer kann an jeder Stelle von ANCA in den Dialog eingreifen; die Bildschirm-Ein-/Ausgabe steht deshalb im Mittelpunkt der Systemsteuerung. Zu Beginn eines "Steuerzyklus" wird aus der Bilddatei das erforderliche Bild abgerufen, (in der Regel) mit Zusatzinformationen versehen und auf dem Bildschirm ausgegeben. Zu diesem Zeitpunkt werden der erreichte Dialogstand sowie spezielle Zustandsvariable und Zwischenergebnisse abgespeichert; damit soll es ermöglicht werden, bei Verarbeitung mit anderen Datenkommunikationssystemen (z. B. IMS), bei denen das Programm an dieser Stelle ausgela-

gert wird, wieder zum erreichten Dialogstand zurück-zukehren¹⁾.

Die Eingabe wird speziell aufbereitet und in einem eigenen Unterprogramm geprüft. Bei Auftreten eines Fehlers wird das letzte Bild zusammen mit einer Fehlermeldung wiederholt; im anderen Fall verzweigt die Programmsteuerung entsprechend der eingegebenen Information und dem vorliegenden Dialogstand.

Handelt es sich bei der Eingabe um ein Betriebssystemkommando, so wird es mit höchster Priorität abgearbeitet, anschließend geht der normale Dialogablauf weiter. Ein Vorrangkommando soll in der Regel den Ablauf verändern; ANCA prüft die Zulässigkeit der Eingabe in der vorliegenden Phase und variiert den Ablauf entsprechend. Bei "normaler" Eingabe wird in den entsprechenden Modul von ANCA verzweigt; dort erfolgt die Feinsteuerung zur Fortführung des Programmablaufs. Der Zyklus schließt sich mit der folgenden Ein-/Ausgabeoperation.

Auf die erforderlichen Dateien, Informationstabellen und Schalter zur Steuerung soll hier nicht eingegangen werden.

4.2.4. Dateikonzept

ANCA benötigt zur Clusteranalyse (maximal) 16 Dateien; entsprechend der Definition von ANCA als selbständiger Teilmenge von MADAS (vgl. Abschnitt 2.3.3.) können ei-

¹⁾ Vgl. dazu Lorenz, N., Ein Programmpaket zur Cluster-Analyse im Rahmen des Dialogsystems MADAS unter besonderer Berücksichtigung der Unterstützung des Benutzers Diplomarbeit, Nürnberg 1977.

nige der Dateien entweder vom Benutzer oder von MADAS erstellt worden sein. Aus Akzeptanzgründen ist sichergestellt, daß ANCA vom Anwender nur die Angabe seiner spezifischen Dateien verlangt; die übrigen Dateien werden automatisch angemeldet.

Im einzelnen benötigt ANCA:

- Eine Datei mit den Originaldaten

Ihr Format ist relativ freigestellt; Vorschriften existieren für die Kennzeichnung der verschiedenen Satzarten. Sie kann von MADAS als Teilmenge der angeschlossenen Datenbank erzeugt werden.

- Eine Datendefinitionsdatei

Sie teilt dem System das Format der Daten sowie Stamminformationen der einzelnen Merkmale (Skalierung, Ausprägungszahl usw.) mit und erlaubt so die Verarbeitung unterschiedlicher Probleme. Sie kann von MADAS bereitgestellt werden.

- Eine Bilddatei

Sie enthält alle Dialogbilder; sie wird automatisch bereitgestellt.

- Eine Ausprägungsdatei (wahlfrei)

Sie enthält die Bezeichnungen der einzelnen Ausprägungsstufen zur Unterstützung der Ergebnisinterpretation. Sie kann Informationen von MADAS enthalten; sie ist nur bei Bedarf bereitzustellen.

- Eine Ergebnisdatei (wahlfrei)

Sie nimmt charakteristische Werte über die verschiedenen Ergebnisse auf und erlaubt so Vergleiche von

Ergebnissen und Analysen einzelner Cluster. Sie ist nur bei Bedarf anzumelden.

- Bis zu zehn Clusterdateien (wahlfrei)

Sie nehmen jeweils die Datensätze auf, die zu einem Cluster gehören. ANCA erzeugt sie automatisch bei Aufruf des entsprechenden Bausteins.

- Eine Plotdatei (wahlfrei)

Sie beinhaltet die Information, die zum Plotten eines Bildes erforderlich ist. Sie wird automatisch angemeldet.

Für den Fall, daß MADAS die Datendefinitionsdatei erstellt, wird vom Benutzer nur die Angabe einer Datei zwingend verlangt.

4.2.5. Schnittstellenkonzept

Aus Gründen der Flexibilität und Erweiterbarkeit des Systems wird großer Wert auf klar definierte Schnittstellen gelegt; im einzelnen seien aufgeführt:

- Schnittstelle zwischen Programmoduln

Jedes Modul besitzt genau einen Ein- und Ausgang; die Steuerparameter sind aufgeführt.

- Schnittstelle Rahmen - Betriebssystem der TR 440

Der Übergang auf die Ebene des Betriebssystems ist nur im zentralen Steuerbereich möglich und entsprechend dokumentiert.

- Schnittstelle Rahmen - Vorrangsystem

Vorrangkommandos werden in einem speziellen Modul abgearbeitet, von dem aus die Steuerung vorgenommen wird.

- Schnittstelle Rahmen - Daten

Die Datendefinitionsdatei stellt die Schnittstelle zu den Daten dar; sie ermöglicht, daß die eigentlichen Verarbeitungsbausteine datenunabhängig arbeiten.

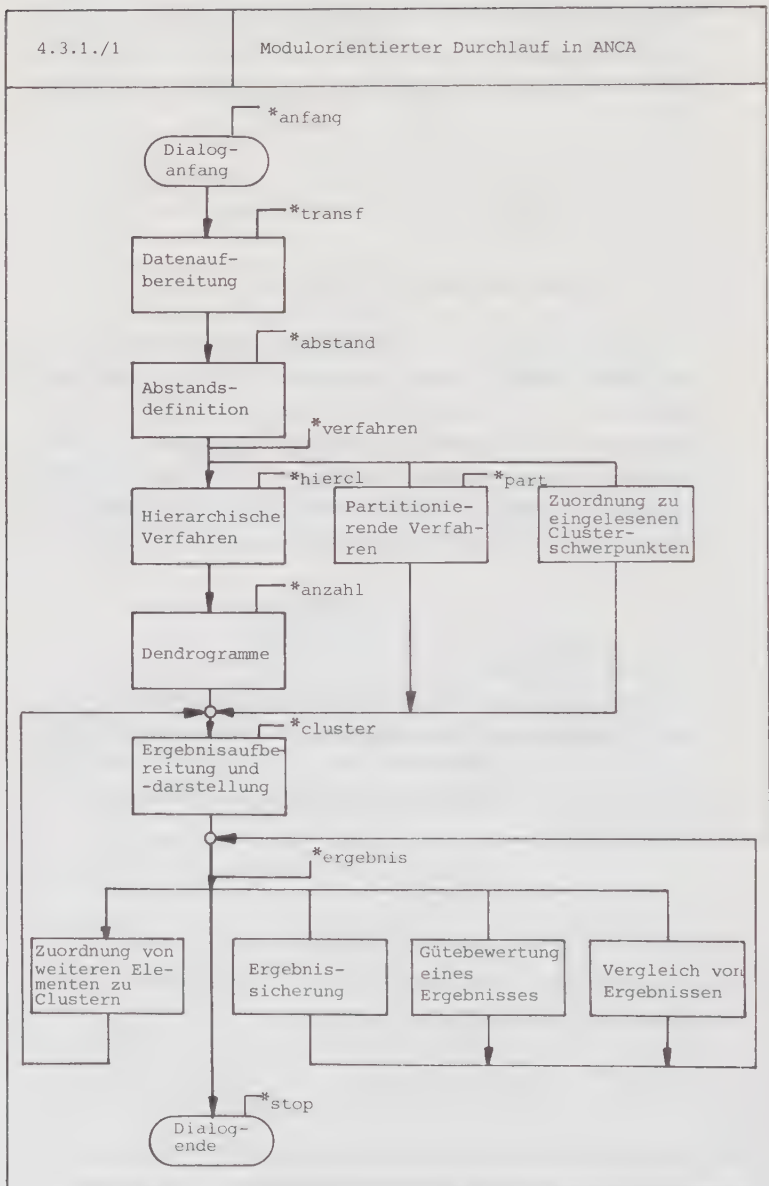
4.3. Dialogablauf

4.3.1. Modulorientierter Durchlauf

Der Ablauf einer Analyse wird vom System entsprechend den Phasen eines Clusterprozesses geführt (vgl. Abbildung 4.3.1./1). Es existiert kein zentraler Steuerblock, der nach jedem Teilschritt erscheint und alle Möglichkeiten des Programmsystems anbietet. Vielmehr werden in Auswahlmenüs die Möglichkeiten dargestellt, die in einer bestimmten Dialogphase sinnvoll sind.

4.3.2. Kommandoorientierte Steuerung

Um dem erfahrenen Benutzer dennoch die Möglichkeit zur Eigensteuerung des Ablaufs zu geben, erlaubt ANCA eine Fülle von Kommandos, die in der Regel Auswirkungen auf den Dialogfortgang haben (vgl. Abbildung 4.3.2./1).



4.3.2./1	Kommandos in ANCA

AUFLISTUNG ALLEER DIALOGKOMMANDOS:	
*ABSTAND	MENUKARTE DER ABSTANDSBESTIMMUNGEN
*ANFANG	AN DEN ANFANG DES DIALOGS ZURUECKKEHREN
*ANZAHL	ANDERE CLUSTERANZAHL WAEHLEN
*AUSWAHL	CLUSTER FUER VARIANZ- ODER DISKRIMINANZANALYSE AUSWAHLEN
*BATCH	WEITERVERARBEITUNG IM BATCHBETRIEB
*CLUSTER	MENUKARTE DER AUSGABE VON CLUSTERKENNDAENEN
*DRUCKAUS	DRUCKERPROTOKOLL AUSSCHALTEN
*DRUCKEIN	DRUCKERPROTOKOLL EINSCHALTEN
*DRUCKER	SOFORTIGE AUSGABE DES BISHER ERSTELLTEN DRUCKERPROTOKOLLS
*ERGEBNIS	MENUKARTE DER ERGEBNISBEHANDLUNG
*HILFE/XXXX	ERKLAERUNG VON BEGRIFFEN
*HIERCL	MENUKARTE DER HIERARCHISCHEN VERFAHREN
*LEXIKON	LISTE ALLER IM LEXIKON GESPEICHERTEN BEGRIFFE
*PART	PARTITIONIERENDES VERFAHREN
*STOP	BEENDIGUNG DES DIALOGS
*TRANSF	MENUKARTE DER DATENAUFBEREITUNGEN
*VERFAHREN	MENUKARTE DER CLUSTERUNGSVERFAHREN
SEND.	
\>	

Beispielsweise ist es im Rahmen der Sensitivitätsanalyse möglich, bei gewünschten Änderungen von Parametern oder Teilmethoden gezielt zurückzuspringen:

- Mit *ANFANG zum Beginn von ANCA
- Mit *TRANSF zur Datenvorbehandlung
- Mit *ABSTAND zur Abstandsbestimmung
- Mit *VERFAHREN zur Wahl der Methodengruppe
- Mit *PART zum partitionierenden Verfahren
- Mit *HIERCL zum hierarchischen Verfahren
- Mit *ANZAHL zur Bestimmung der Clusterzahl
- Mit *CLUSTER zum Beginn der Interpretationsroutine für Cluster

Als mächtigstes Kommando für Experten existiert ein Schnellgang. Mit einer einzigen Eingabe im Eröffnungsbild können alle Teilmethoden benannt werden; der Dialog wird erst mit den Interpretationsbausteinen nach Durchrechnung der Analyse fortgesetzt.

Dagegen ist keine generelle Unterscheidung von Experten- und Laienmodus mit unterschiedlich ausführlicher Kommunikation vorgesehen. Der Dialog ist insgesamt ausführlich gehalten; für Laien ist zudem eine detaillierte Beschreibung einzelner Phasen über HILFE-Bilder möglich.

5. Kritische Analyse der Entwicklung und Anregungen für zukünftige Aktivitäten

5.1. Diskussion des Gesamtkonzepts

5.1.1. Zielgruppendifkussion

Ziel der Entwicklung von ANCA war nicht primär, die technische Leistung "Clusteranalyse" für alle Anwendungsbereiche mit einer umfangreichen Methodensammlung, wie sie z. B. CLUSTAN besitzt, zu erbringen; im Vordergrund stand vielmehr, die Clusteranalyse mit Bausteinen zur Verbesserung der Akzeptanz zu versehen und zu testen, ob sich damit neue Anwender gewinnen lassen, die mit konventionellen Programmsystemen nicht oder nur mit Schwierigkeiten arbeiten.

Eine schlüssige Aussage über den Grad der Zielerreichung ist zur Zeit noch nicht möglich. Erfreuliche Aspekte ergaben sich in wissenschaftlicher Umgebung. Hier wurden Forschungsarbeiten - zum großen Teil Diplomarbeiten - unterstützt. Die Bearbeiter entsprachen der in Abschnitt 2.2.1. definierten Zielgruppe; ihre Kenntnisse über Clusteranalyse gingen über das durch Lehrveranstaltungen vermittelte Grundwissen nicht hinaus. Dennoch wurde eine Reihe von befriedigenden Ergebnissen erzielt. Besonderes Interesse fanden die Interpretationsbausteine und die Möglichkeit, die Cluster einzeln zu analysieren; Probleme traten vor allem bei der Auswahl der geeigneten Methode auf.

Von Seiten der Unternehmenspraxis gab es Vorbehalte gegenüber einer Anwendung. Meist wird die Clusteranalyse zum Instrument für Methodenspezialisten erklärt; obwohl Interesse für Analysen besteht, verzichtet man

häufig ganz auf eine Anwendung oder vergibt das Problem an Externe. Gespräche bestätigen oft, daß die Clusteranalyse selbst Ursache für dieses Verhalten ist. Ihre bereits angesprochene "Benutzerunfreundlichkeit" läßt sie dem Anwender gegenüber komplizierter erscheinen als sie wirklich ist; entsprechend kann sie die Barriere nicht überwinden, die sich allgemein gegenüber neuen und anspruchsvollen Verfahren ergibt¹⁾.

Die häufig angesprochene Einschaltung von Methodenspezialisten²⁾ verspricht ebenfalls keine besseren Zukunftsaussichten. Die Vielzahl der zur Verfügung stehenden Methoden und die Notwendigkeit, mit verschiedenen Methoden und Paketen zu arbeiten, machen Unterstützung des Anwenders erforderlich. Dies gilt insbesondere für den Typ von Analytiker, der in Zukunft wohl verstärkt zu finden sein wird: Er kennt seine Daten ganz genau und läßt dieses Wissen auch in seine Arbeiten einfließen³⁾. Die Methoden sind für ihn nur Mittel zum Zweck, er beherrscht sie soweit, daß er eine korrekte Anwendung durchführen kann.

1) Zu dieser und anderen neueren Erkenntnissen der Akzeptanzforschung siehe Mertens, P., Computergestützte Unternehmensplanung, Informatik Spektrum 2 (1979), S. 86 ff., insbes. S. 94.

2) Zur allgemeinen Diskussion über Methodenspezialisten vgl. z. B. Winand, U., Zur Gestaltung von Modell-Methoden-Verbund-Programmen im Rahmen von quantitativen modellgestützten Problemlösungen, Angewandte Informatik 17 (1975), S. 503 ff., der für einen solchen Spezialisten eintritt, sowie Möllhoff, L., a.a.O., S. 154 ff. mit dem Befund, daß erste Eindrücke eher zu Pessimismus raten.

3) Dieser Typ von Benutzer ist für Clusteranalysen sehr wichtig; vgl. z. B. Dubes/Jain: "Someone involved in the subject matter of the data base is usually able to interpret the results of a cluster analysis more meaningfully than the clustering professional." Dubes, R. und Jain, A. K., Clustering Techniques: The User's Dilemma, Pattern Recognition 8 (1976), S. 247 ff., hier S. 258.

Insgesamt ergibt sich, daß zum jetzigen Zeitpunkt ein anwenderorientiertes System zur Clusteranalyse im Bereich der Hochschule positive Aufnahme findet, in den Unternehmen bestehen dagegen noch Vorbehalte. Weitere Forschungsarbeiten - z. B. im Zusammenhang mit den angesprochenen Problemen der Methodenwahl - sind erforderlich; die Generation von Studenten, die jetzt in der Hochschule ein Grundwissen über multivariate Verfahren erhält, sollte dann in der Lage sein, die notwendigen praktischen Erfahrungsberichte zur Akzeptanz fortgeschrittener Methodenbankansätze zu liefern.

5.1.2. Diskussion der interaktiven Verarbeitung

ANCA hat in den Fallstudien die Erwartungen in bezug auf Schnelligkeit der Analyse, Steuerung des Ablaufs und Beurteilung der Lösung im Dialog erfüllt; insbesondere die Darstellungsformen der Ergebnisse wurden positiv anerkannt.

Dagegen haben die Einschränkungen für die Problemgrößen gelegentlich Schwierigkeiten bereitet.

Die Zahl der Objekte ist dabei - trotz der von Fachleuten zum Teil genannten Zahlen¹⁾ - relativ unkritisch. Wie auch in der Literatur anerkannt²⁾, tritt nach etwa 100 Objekten eine Stabilisierung des Ergebnisses ein, die eine Zuordnung der übrigen Daten ermöglicht. Mit ANCA konnte diese Erkenntnis bestätigt werden; durch die Zuordnung wurde nur noch in seltenen Fällen eine Veränderung der Schwerpunkte erzielt.

¹⁾ Wendt spricht von bis zu 4.000 Objekten, Späth von 50 bis 5.000. Wendt, F., a.a.O., S. 76; Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O., S. 12. In fünf von elf der bei Späth aufgeführten Fallstudien wurden dagegen weniger als 100 Sätze klassifiziert; vgl. Späth, H. (Hrsg.), Fallstudien ..., a.a.O.

²⁾ Vgl. z. B. Wendt, F., a.a.O., S. 62; Landgrebe, K. P., Konstruktionsprinzipien und Erkenntnisgehalte neuerer Bevölkerungsmodelle für das Marketing, in: Braunschweig, E. (Hrsg.), a.a.O., S. 150.

Die Zahl der Cluster hält sich ebenfalls im Rahmen dessen, was von Anwendern gewünscht wird¹⁾; gelegentlich tritt dennoch der Wunsch nach einer höheren Gruppenzahl auf.

Kritisch ist dagegen die Zahl der Merkmale. Zwar gefährden bei einer Vielzahl von Variablen Kompensationswirkungen die Eindeutigkeit und Interpretierbarkeit eines Ergebnisses. Die Zahl von sieben Merkmalen ist aber in der Regel dennoch zu niedrig²⁾. Als Ausweg bietet es sich an, Analysen mit wechselnden Merkmalen durchzuführen und die Ergebnisse zu vergleichen; diese Vorgehensweise wird aber nicht immer angenommen.

Eine Erweiterung der Zahl der Variablen ist von der internen Verarbeitungskapazität und der Definition der Dateien her leicht möglich; bei Bedarf kann noch auf eine Variante zurückgegriffen werden, bei der die Clustermatrix auf Platte ausgelagert wird³⁾. Erforderlich wären aber Bildschirme mit größerer Darstellungskapazität. Die Aufteilung der Informationen auf mehrere Bildschirme wurde aus Akzeptanzgründen nicht in Erwägung gezogen.

1) In acht der Späth'schen Fallstudien wurden weniger als zehn Cluster gewählt; vgl. Späth, H. (Hrsg.), Fallstudien ..., a.a.O.

2) Vgl. ebenda; nur in einem Fall wurden weniger als sieben Merkmale verwendet; in der Regel lag die Zahl bei 30 Merkmalen. Vgl. dagegen die Angaben von zwei bis zehn Variablen bei Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O., S. 12.

3) Vgl. z. B. für hierarchische Verfahren: Molzberger, P. und Peschanel, F. D., Entwicklung schneller Verfahren zur hierarchischen Cluster-Analyse bei großen Datenmengen, in: Kohlas, I., Seifert, O., Stähly, P. und Zimmermann, H. J. (Hrsg.), Proceedings in Operations Research 5, Würzburg 1975, S. 187 ff. Für partitionierende Verfahren: Späth, H., Partitionierende Cluster-Analyse für große Objektmenngen mit binären Merkmalen am Beispiel von Firmen und deren Berufsgruppenbedarf, in: Späth, H. (Hrsg.), Fallstudien ..., a.a.O., S. 6) ff.

Insgesamt lassen sich aber auch bei unveränderter Lösungskapazität ausreichend Anwendungsfälle für ANCA finden. Der Einsatz der interaktiven Verarbeitung ist ohne Einschränkung zu empfehlen, wenn eine Datenmenge neu zu analysieren ist; bei Wiederholungsstudien, für die der Ablauf fest vorgegeben ist, kann dagegen auch in Batch-Verarbeitung übergeben werden. Hier lassen sich Parallelen zur Prognoserechnung finden, in der die Analyse und Modellbildung sinnvollerweise im Mensch-Maschine-Dialog durchgeführt wird, die routinemäßige Prognose mit Stapelverarbeitung.

Als Testfall für die Stapelverarbeitung in ANCA wurde das Programm CLUDIB realisiert (vgl. Kapitel 3.1.1.1.). Eine generelle Übergabeschnittstelle für alle Verfahren existiert nicht; sie kann mit Betriebssystemkommandos (und somit mit eingeschränkter Portabilität) geschaffen werden. Es bieten sich zwei grundsätzliche Varianten der Batchverarbeitung an:

- Nach der Übergabe in die Stapelverarbeitung werden die Ergebnisse auf Liste ausgegeben; der Dialog dient nur noch zur fehlerfreien Kombination der einzelnen Bausteine zur Methode.
- Es wird nur für die eigentliche Rechenzeit in die Stapelverarbeitung übergeben; die Ergebnisse werden aufgefangen und können anschließend mit den üblichen Interpretationsbausteinen analysiert werden.

Variante 1 hat den Vorteil, daß größere Objekt- und Merkmalszahlen verarbeitet werden können, da die Restriktionen des Bildschirms weitgehend wegfallen; andererseits entfallen einige der Akzeptanzbausteine, die eine Methodenbank ausmachen.

Variante 2 bietet weiterhin allen Komfort der Methodenbank, erlaubt aber die Verwendung aufwendigerer Algorithmen und größerer Datenmengen. Nach der Initialisierung der Methode kann der Bildschirm verlassen werden; die Ergebnisinterpretation erfolgt zu einem beliebigen Zeitpunkt, eventuell unterstützt durch eine Liste, die in der Zwischenzeit gedruckt wurde.

Insgesamt erscheint Variante 2 empfehlenswerter; bei Bedarf können beide implementiert werden, wobei Variante 1 auf große Datenmengen zu beschränken ist.

5.1.3. Möglichkeiten Graphischer Datenverarbeitung

Ein Grund für den erfolgreichen Einsatz von ANCA in Fallstudien ist die Verwendung vollgraphischer Bildschirme bzw. Plotter bei der Darstellung des Dendrogramms und der Profile der Cluster. Eine Ausweitung der Einsatzmöglichkeiten scheint empfehlenswert.

Systeme zur Clusteranalyse (z. B. CLUSTAN) produzieren neben Dendrogrammen zum Teil noch zweidimensionale Streumatrizen, die die Verteilung aller Objekte in bezug auf zwei Merkmale oder Faktoren zeigen. Diese Form wurde nicht übernommen. Erfolgversprechender scheint der Versuch, mehr als zwei Dimensionen auf der Ebene darzustellen. Damit ist auch der Ansatz des

"Visual Clustering"¹⁾ realisierbar.

Die nächste Stufe könnte die Eingabe von Informationen sein, z. B. das Einzeichnen von Cluster-Schwerpunkten oder das Ausschalten von Außenseitern aus einer Analyse.

Erforderlich sind Verbesserungen im Zugang zur Graphik-Software, wie es etwa in graphischen Methodenbanken²⁾ versucht wird. Die Experimentierphase für Graphische Datenverarbeitung ist zwar noch nicht überwunden³⁾, der Preisverfall der betreffenden Hardware läßt aber deutliche Fortschritte für die nächste Zeit erwarten.

5.2. Diskussion des Methodenbereichs

5.2.1. Methodensammlung

Die in Abschnitt 3.1.1.1. angesprochene Kritik an der vorliegenden Methodensammlung ist berechtigt; tatsächlich ist nur ein kleiner Teil der Verfahren implementiert, die einem Spezialisten der Clusteranalyse zur Verfügung stehen. Andererseits hat sich bei der Anwendung von ANCA bestätigt, daß eher ein Bedürfnis nach wenigen Methoden besteht, deren Einsatzspektrum dafür klar umrissen ist. Auch bei den vorhandenen Methoden

¹⁾ Kamen, J., Quick Clustering, Journal of Marketing Research 7 (1970), S. 199 ff.

²⁾ Vgl. Encarnacao, J., Konzepte, Realisierung und Erfahrungen mit dem Einsatz graphischer Informationssysteme, Vortrag in Erlangen am 27. 2. 1978, und Heinen, H. J., und Konkart, R., GRADAS - ein interaktives grafisches Informationssystem auf der Grundlage einer relationalen Datenbank, Interner Bericht der AEG, o. O., o. J.

³⁾ Wißkirchen, P., Stand und Entwicklungstendenzen der Grafischen Datenverarbeitung, Interner Bericht der GMD, St. Augustin 1978.

gab es immer wieder Auswahlprobleme, die nicht leicht zu lösen waren; dies betraf auch das Zusammenwirken der einzelnen Bausteine.

Langfristiges Ziel muß es sein, in die Methodensammlung Verfahren aufzunehmen, die Abstandsmessung, gegebenenfalls Standardisierung und die eigentliche Clusterbildung umfassen. Dies entspricht auch der gegenseitigen Abhängigkeit der einzelnen Komponenten.

Kurzfristig wären höchstens kleinere Ergänzungen des Methodenspektrums sinnvoll. Dabei ist vor allem daran zu denken, neben KMEANS ein weiteres partitionierendes Verfahren einzubeziehen, das auch für nichtmetrische Daten definiert ist. Ferner sollte der Einsatz eines hierarchischen Verfahrens, das auf der Basis der Entropieanalyse¹⁾ arbeitet, geprüft werden; Vogel hält dieses Vorgehen speziell bei binären Merkmalen für sinnvoll²⁾. Bei allen Erweiterungen um Clusteralgorithmen ist aber zu bedenken, daß in der Regel mehrere Läufe mit einfacheren Methoden und ein Vergleich der Ergebnisse sinnvoller sind als die Anwendung einer komplizierten, die mindestens genausoviel kostet³⁾.

Bei den Abstandsmaßen bietet es sich an, ein Verfahren auf der Basis des Korrelationskoeffizienten zu implementieren. Zwar gibt es dagegen viele Einwände⁴⁾; ande-

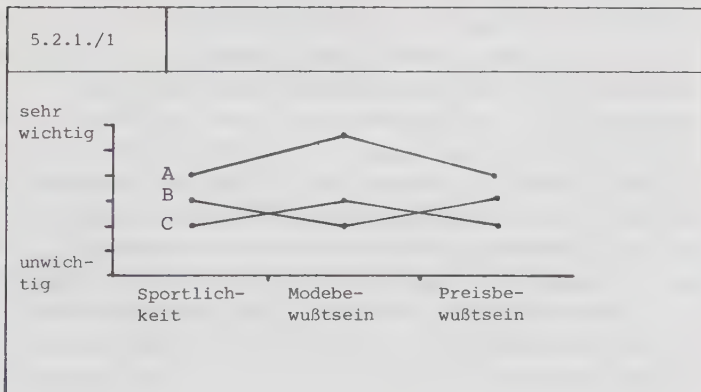
¹⁾ Vgl. Vogel, F., Probleme ..., a.a.O., S. 252 ff.

²⁾ Vgl. ebenda, S. 335.

³⁾ Vgl. z. B. Dubes, R. und Jain, A. K., Clustering ..., a.a.O., S. 259.

⁴⁾ Vgl. dazu das Zitat von Colless bei Vogel: "It is an unfortunate historical accident that pioneer studies in numerical taxonomy often employed, as an estimate of resemblance, the product-moment correlation (r)"; Vogel, F., Probleme ..., a.a.O., S. 92 sowie Bock, H. H., Automatische Klassifikation, a.a.O., S. 46.

rerseits ist es mit keinem anderen Maß möglich, ähnliche Merkmalsstrukturen auf unterschiedlichem Niveau auch als ähnlich zu erkennen. Anwendungsbeispiel hierfür ist die Vergabe von Noten durch mehrere Gutachter, die von einem unterschiedlichen Beurteilungslevel ausgehen; dann ist nicht die Bewertung per se entscheidend, sondern die Relation der Beurteilungen zueinander; ähnlich ist es bei der Einschätzung von Eigenschaften als "wichtig" oder "weniger wichtig" (vgl. Abbildung 5.2.1./1): Wenn die Profile von A und C trotz der unterschiedlichen Höhe als ähnlich erkannt werden sollen, ist ein Korrelationsmaß zu wählen.



Dagegen erscheint eine Einbeziehung des häufig empfohlenen MAHALANOBIS-Maßes nicht sinnvoll; es ist stark umstritten¹⁾ und genügt somit nicht den in Abschnitt 3.1.1.1. gestellten Bedingungen.

¹⁾ Vgl. z. B. Bock, H. H., Automatische Klassifikation, a.a.O., S. 40 ff. und Vogel, F., Probleme ..., a.a.O., S. 90 ff.

Insgesamt bleiben die Vorbehalte gegenüber einer Ausdehnung der Methodensammlung bestehen. Bei jeder Erweiterung sind die Charakteristika der einzelnen Verfahren dem Anwender gegenüber deutlich zu machen; nur so kann er sinnvollen Gebrauch davon machen.

5.2.2. Methodenwahl

5.2.2.1. Forderung nach einer verbesserten Methodenwahl

Beim jetzigen Ausbaustand der Methodensammlung sind verbesserte Auswahlhilfen zur Methodenwahl wünschenswert; bei der von Statistikern geforderten Erweiterung der Sammlung sind sie unerlässlich. Für eine sinnvolle Selektion sind neben den in Abschnitt 3.2.3. aufgeführten Kriterien weitere Eigenschaften der zu analysierenden Daten und Vorstellungen über die Aufteilung der Objekte auf Klassen auszuwählen und - im Idealfall - in einen Prozeß der Mustererkennung einzubetten. Dabei kann der Computer nicht zum "kompetentesten Datenanalytiker" werden; diese Vorstellung ist "von aller Realität weit entfernt"¹⁾. Vielmehr geht es darum, in einem interaktiven Prozeß die Vorteile von Mensch und Maschine zu verknüpfen und somit den Auswahlvorgang zu unterstützen, nicht automatisch abzuwickeln.

¹⁾ Vgl. die Diskussion hierzu bei Allerbeck. Er leitet aus dieser Tatsache ab, daß jedes Verbiegen von falschen Analysetechniken durch den Computer reine Utopie ist, ohne den Weg des Zusammenwirkens von Mensch und Maschine zu sehen. Andererseits verlangt er zu Recht eine verstärkte Ausbildung im Gebrauch von Datenanalysesystemen. Siehe Allerbeck, K. R., Datenanalyse, in: Koolwijk, J. van und Wieken-Mayser, M. (Hrsg.), Techniken der empirischen Sozialforschung, München 1974, S. 185 f.

Die relevanten Kriterien zur Methodenwahl können im wesentlichen aus folgenden Quellen abgeleitet werden:

1. Wissenschaftler beschreiben die Voraussetzungen einzelner Verfahren, die von ihnen entwickelt oder geprüft wurden
2. Wissenschaftler vergleichen mehrere Verfahren mit echten oder gezielt ausgewählten synthetischen Daten
3. Praktiker berichten über ihre Erfahrungen bei der Bearbeitung von Problemen mit Verfahren der Clusteranalyse

Voraussetzung für eine Verwendung der Kriterien ist, daß sie entweder dem Anwender direkt zugänglich sind oder maschinell geprüft werden können. Diese Bedingung erfüllen insbesondere Hinweise der Kategorie eins häufig nicht, weil nur die theoretischen - z. B. mathematischen oder statistischen - Grundlagen der Verfahren beschrieben werden¹⁾. Vergleiche von Verfahren versprechen grundsätzlich die größten Hilfen; sie werden deshalb im folgenden Abschnitt gesondert aufgeführt. Bei Erfahrungsberichten von Praktikern werden selten die Beweggründe für die Auswahl der einzelnen Methoden genannt²⁾, zum Teil finden sich widersprüchliche Vorgehensweisen³⁾.

1) Vgl. die Kontroverse der pragmatisch vorgehenden "Australischen Schule" mit der "Cambridge-Schule", dokumentiert etwa in Williams, W. T., Lance, G. N., Dale, M. B. und Clifford, H. T., *Controversy concerning the Criteria for taxonomic Strategies*, *Computer Journal* 14 (1971), S. 162 ff.

2) Eine Zusammenstellung von Hinweisen aus Fallstudien findet sich in Leisch, E., *Bestandsaufnahme ...*, a.a.O., insbes. S. 76 ff.

3) Vgl. z. B. die unterschiedliche Behandlung ordinaler Daten bei den verschiedenen Autoren des von Späth herausgegebenen Buches: Späth, H. (Hrsg.), *Fallstudien ...*, a.a.O., S. 16, S. 162 und S. 183.

In Abschnitt 5.2.2.3.1. werden Kriterien zusammengestellt, die bei einer anwenderbezogenen Methodenwahl Verwendung finden können. Es handelt sich dabei gewissermaßen um einen Maximalkatalog, der aus allen genannten Quellen abgeleitet wurde. Er gilt unter der Voraussetzung, daß eine wesentlich erweiterte Methodensammlung die Akzeptanzschranke überwunden hat, und ist deshalb weitgehend als spekulativ zu bezeichnen¹⁾.

In Abschnitt 5.2.2.3.2. folgen Überlegungen zu einem Mustererkennungsansatz, der über einen reinen Schablonenvergleich hinausgeht. Sie beschränken sich auf einige grundlegende Gedanken; eine Realisierung des Ansatzes setzt stärkere Übereinstimmung von Kriterienkatalog und Methodenangebot voraus.

5.2.2.2. Vergleiche von Verfahren zur Clusteranalyse in der Literatur

Vergleiche von Verfahren zur Clusteranalyse existieren in relativ geringer Zahl in der Literatur. Offensichtlich bereitet sowohl die Zahl der relevanten Verfahren und Parameter als auch die Fülle verschiedener Datenkonstellationen erhebliche Schwierigkeiten; darüber hinaus fehlt ein objektiver Maßstab für die Güte der gefundenen Lösungen.

Einige Autoren stellen offensichtlich die Methodologie beim Verfahrensvergleich in den Vordergrund ihrer Betrachtung. So definiert Rand²⁾ ein Maß für die Beurteilung verschiedener Clusteralgorithmen sowie vier grund-

1) Vgl. auch Dubes und Jain: "Little agreement exists on appropriate criteria", in: Dubes, R. und Jain, A. K., Models ..., a.a.O., S. 153.

2) Rand, W. M., Objective Criteria for the Evaluation of Clustering Methods, Journal of the American Statistical Association 66 (1971), S. 846 ff.

sätzliche Prüfhypothesen, um dann eher beispielhaft zwei Verfahren zu vergleichen. Green und Rao¹⁾ untersuchen zehn Abstandsmaße zur Clusteranalyse; im wesentlichen bestimmen sie aber nur die Ähnlichkeit der verschiedenen Varianten und stellen das Ergebnis - recht originell - mit Hilfe der mehrdimensionalen Skalierung dar; echte Auswahlhilfen liefern aber auch sie nicht²⁾.

5.2.2.2.1. Vergleiche von hierarchischen Verfahren

Die meisten Verfahrensvergleiche existieren für hierarchische Verfahren. Eine der meistzitierten Untersuchungen stammt von Cunningham und Ogilvie³⁾; die Autoren definieren sechs zweidimensionale Datenstrukturen, die teilweise mit "noise" überlagert sind. Zwei der von ihnen gefundenen Ergebnisse finden sich nahezu in allen späteren Vergleichen: Das Verfahren nach Ward (entspricht HIERCL 7 in der Sammlung von ANCA, vgl. dazu Abschnitt 3.1.1.1.) findet "richtige" Ergebnisse bei etwa gleich großen Clustern, bevorzugt auch bei metrischem Skalenniveau; andererseits versagt es bei verketteten Daten. Ein genau gegensätzlicher Befund ergibt sich für das Verfahren Single Linkage (entspricht HIERCL 1); es erkennt bevorzugt verkettete Strukturen, kann aber für metrische Daten selten ein vernünftiges Ergebnis liefern. Im übrigen erweist sich in den Tests das Verfahren Group Average (HIERCL 5)

1) Green, P. E. und Rao, V. R., A Note on Proximity Measures and Cluster Analysis, Journal of Marketing Research 6 (1969), S. 359 ff.

2) Die Autoren beklagen am Ende ihrer Ausführungen das Fehlen einer festen Basis für vergleichende Untersuchungen und betonen ihre Notwendigkeit, gerade in Anbetracht der zunehmenden Zahl von Clusterprogrammen; vgl. ebenda, S. 364.

3) Cunningham, K. M. und Ogilvie, J. C., Evaluation of hierarchical Grouping Techniques: a preliminary Study, Computer Journal 15 (1972), S. 209 ff.

als bestes; die Unterschiede zu den anderen Methoden (bis auf HIERCL1) sind unerheblich.

Im Gegensatz hierzu ergibt sich bei Mojena¹⁾ eine klare Überlegenheit des Verfahrens nach Ward, was seiner Meinung nach wegen der Auswahl streng metrischer - künstlicher - Daten eine "self-fulfilling prophecy"²⁾ darstellt; die Verfahren 2, 5 und 6 liefern ausgesprochen schlechte Ergebnisse, 1 zeigt wieder den typischen Ketteneffekt.

Forst³⁾ führt zwei Vergleichsarten durch: einen Versuch mit Daten bekannter Gruppenzugehörigkeit, bei denen Fehlklassifizierungen gemessen werden können, und einen mit Zufallsdaten, allerdings eingeschränkt auf normalverteilte zweidimensionale Daten. Zusätzlich zu den hierarchischen Verfahren bezieht er das partitionierende KMEANS in die Analyse ein. Forst stellt fest, daß von einer generellen Unterlegenheit hierarchischer Verfahren keineswegs gesprochen werden kann; insbesondere scheinen sie schwächer auf Verletzungen eines zugrundeliegenden theoretischen Modells zu reagieren. Interessant ist seine Analyse über die Reaktion der Methoden auf Außenseiter: KMEANS und das entsprechende hierarchische Verfahren nach Ward nehmen sie in größere Gruppen auf und stellen - bei verminderter Homogenität - die dominierende Gruppenstruktur dar; HIERCL 1, 3, 4 und 6 isolieren dagegen Außenseiter (und reagieren somit empfindlich auf Zufallsabweichungen).

¹⁾ Mojena, R., Hierarchical Grouping Methods and Stopping Rules: An Evaluation, Computer Journal 20 (1977), S. 359 ff.

²⁾ Ebenda, S. 362.

³⁾ Forst, H. T., Zur Klassifizierung ..., a.a.O., S. 45 ff.

Eine Reihe von Autoren rechnen jeweils ein kleines Beispiel mit allen sieben Verfahren durch¹⁾; daraus leitet dann beispielsweise Späth Empfehlungen für HIERCL 6 und 7 bei metrischen (mit EUKLID berechneten) Abständen und für HIERCL 3, 5 und 2 (in dieser Reihenfolge) bei beliebigen Daten ab. Wie auch andere Autoren warnt er bei HIERCL 1, von HIERCL 4 rät er ab²⁾.

Hinweise auf weitere Verfahrensvergleiche gibt Bock³⁾. Ein Beispiel aus der Mustererkennungsliteratur - mit weiteren Quellenangaben - findet sich bei Dubes und Jain⁴⁾.

Insgesamt läßt sich feststellen, daß neben der speziellen Eignung von HIERCL 1 für gekrümmte, etwa bananenförmige Cluster und der Empfehlung für HIERCL 7 bei runden, etwa gleich großen Gruppen wenig übereinstimmende Ergebnisse erzielt werden; dies mag an dem "Ad-hoc-Charakter" vieler Untersuchungen liegen.

5.2.2.2. Sonstige Vergleiche

Vergleiche von hierarchischen mit nicht-hierarchischen Verfahren finden sich selten in der Literatur; so beruft sich etwa Späth auf seine Erfahrungen⁵⁾, wenn er nicht-hierarchische Methoden bevorzugt, ohne ein konkretes Testbeispiel vorzugeben. Eine kritischere Mei-

1) Vgl. z. B. Vogel, F., Probleme ..., a.a.O., S. 249 ff.; Steinhausen, D. und Langer, K., Clusteranalyse ..., a.a.O., S. 82 ff.; Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O., S. 170 ff.

2) Ebenda, S. 189.

3) Bock, H. H., Automatische Klassifikation, a.a.O., S. 400.

4) Dubes, R. und Jain, A. K., Clustering ..., a.a.O.

5) Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O., S. 189.

nung gegenüber hierarchischen Verfahren als Forst (vgl. oben) vertreten Degens und Federkiel¹⁾; in einer Analyse mit synthetischen Daten stellen sie die relativ starke Abhängigkeit hierarchischer Methoden von zufällig verteilten Elementen fest, während nicht-hierarchische Verfahren eher die realistische Datenstruktur finden. In den anderen Darstellungen werden beide Gruppen als Ergänzung gesehen, wobei die Auswahl je nach Zielsetzung zu erfolgen hat.

So untersucht Everitt²⁾ in seiner Studie zwar hierarchische und nicht-hierarchische Verfahren, wobei er auch Methoden einbezieht, die auf der Analyse von Punktedichten beruhen; er vergleicht aber die Ergebnisse für die Gruppen nicht miteinander. Von besonderem Interesse ist die optisch eindrucksvolle Tendenz von Verfahren wie KMEANS, runde Cluster auch dann zu erzeugen, wenn die Daten selbst eigentlich eine andere Einteilung erwarten lassen (vgl. Abbildung 5.2.2.2.2./1)³⁾.

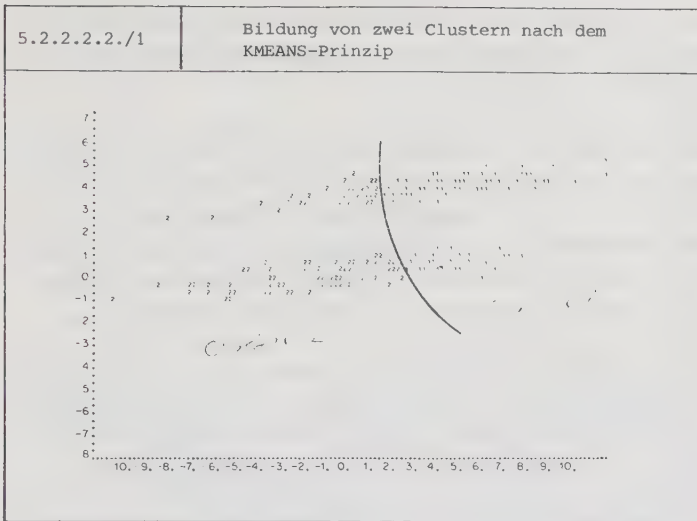
Insgesamt finden sich relativ wenige Vergleiche für partitionierende Verfahren; sie sind zudem häufig nicht darauf ausgerichtet, Entscheidungshilfen für einzelne Methoden in einer bestimmten Konstellation zu geben, sondern den Einsatz beschränkter Ressourcen zu optimieren. So empfiehlt Späth in einer vergleichenden Studie mit zweidimensionalen Testdaten⁴⁾, kein einfacheres Ver-

1) Degens, P. und Federkiel, H., A Monte Carlo Study on Agglomerative Large Sample Clustering, in: Corsten, L.C.A. und Hermans, J. (Hrsg.), COMPSTAT 1978 ..., a.a.O., S. 246 ff.

2) Everitt, B., Cluster Analysis, London u.a. 1974.

3) Ebenda, S. 74.

4) Späth, H., Numerische Erfahrungen zu heuristischen Lösungsverfahren beim Varianzkriterium in der Cluster-Analyse, Angewandte Informatik 19 (1977), S. 67 ff.



fahren als KMEANS (vgl. Abschnitt 3.1.1.1.) zu verwenden, bei begrenztem Budget eventuell nur mit der Standard-Anfangszuordnung (entspricht Anfangszuordnung 1 in Abbildung 3.1.1.1./4); in einer weiteren Studie¹⁾ demonstriert er an denselben Daten, daß eine aufwendigere Analyse bessere Ergebnisse liefern kann, allerdings mit zum Teil zehn- bis zwanzigfachem Rechenaufwand²⁾. Diese Experimente gelten jeweils nur für metrische Daten; eine abschließende Auswahl eines Verfahrens nimmt Späth nicht vor. Die Überlegenheit der Standard-Anfangszuordnung kann im übrigen von Steinhausen und

¹⁾ Späth, H., Computational Experiences with the Exchange Method, European Journal of Operational Research 1 (1977), S. 23 ff.

²⁾ Vgl. Späth, H., Cluster-Analyse-Algorithmen ..., a.a.O., S. 129.

Langer nicht bestätigt werden¹⁾; bei ihren Experimenten ergaben sich Vorteile für eine Anfangspartition, die mit HIERCL 7 erzeugt wurde.

Obwohl der Einfluß der Abstandsmaße auf Ergebnisse der Clusteranalyse unbestritten ist²⁾, existieren - neben theoretischen Gegenüberstellungen der Eigenschaften einzelner Verfahren³⁾ - keine Experimente, die Auswirkungen der Auswahl verschiedener Algorithmen ausweisen; zum Teil werden Tests sogar als nutzlos abgelehnt⁴⁾.

Verfahrensvergleiche insgesamt bringen somit wenig mehr gesicherte Hinweise als in Abschnitt 3.2.3. bereits enthalten sind; weitere Forschungsarbeit scheint hier erforderlich. Dies wird auch von anwendungsorientierten Spezialisten der Clusteranalyse gesehen⁵⁾.

5.2.2.3. Methodenwahl als Mustererkennungsansatz

5.2.2.3.1. Kriterienkatalog

Der im folgenden angeführte Merkmalskatalog verwendet ausschließlich Kriterien, die der Anwender kontrollieren kann.

¹⁾ Vgl. Steinhausen, D. und Langer, K., Clusteranalyse ..., a.a.O., S. 125 f.

²⁾ Vgl. z. B. die oben zitierten Green und Rao: "clustering results are not invariant over these alternative measures"; Green, P. E. und Rao, V. R., A Note ..., a.a.O., S. 359.

³⁾ Vgl. dazu neben den Standardlehrbüchern zur Clusteranalyse z. B. Morrison, D. G., a.a.O.

⁴⁾ Vgl. z. B. Peschanel, F. D., Zu den Konsequenzen der Vernachlässigung von Korrelationen zwischen den Merkmalen bei der Anwendung von Clusteranalysen, in: Dathe, H. N., Mertens, P., Peschanel, F. D., Späth, H. und Zimmermann, H. J. (Hrsg.), Proceedings in Operations Research 6 (1976), S. 224 ff., hier S. 227.

⁵⁾ Beispielsweise wurde diese Forderung auf der Jahrestagung 1979 der Gesellschaft für Klassifikation erhoben. Degens sprach in diesem Zusammenhang von einer notwendigen "Clusteranalyse für Clusteranalyseverfahren".

Die Verbindung mit solchen Merkmalen, die sich ausschließlich aus den Eigenschaften der Verfahren ableiten, findet im nächsten Abschnitt statt.

Die Kriterien lassen sich grob in drei Kategorien einteilen:

- Zielvorstellungen des Analytikers
- Eigenschaften der verwendeten Daten
- Externe Faktoren.

Wichtigste Kategorie für einen benutzerorientierten Ansatz zur Methodenwahl sind die Zielvorstellungen des Anwenders; folgende Festlegungen sind zu treffen:

- Analyse von Objekten oder Merkmalen?

Der Anwender muß angeben, ob er Objekte, Merkmale oder beide in einem Durchgang klassifizieren will; im Normalfall wird er sich für eine Analyse von Objekten entscheiden.

- Suche oder Bestätigung einer Klassenstruktur?

Zur Beschreibung einer vorgegebenen Klassenstruktur können entweder nur die Zahl der Cluster oder zusätzlich deren Schwerpunkte bekannt sein; im ersten Fall erscheint ein partitionierendes Verfahren sinnvoll, im zweiten ein Zuordnungsverfahren. Liegt keine Information vor, empfiehlt sich die Wahl eines hierarchischen Verfahrens, das eine grobe Übersicht liefern kann (vgl. Abbildung 3.2.3.4./2).

- Strukturanalyse oder Außenseitersuche?

Je nach Anwendungsfall kann es sinnvoll sein, entweder eine grobe Strukturierung der Datenmenge vorzu-

nehmen, die den Merkmalsraum etwa gleich aufteilt, oder bewußt Außenseiter - beispielsweise besonders unzufriedene Kunden - herauszufiltern; die Verfahren eignen sich für die beiden Aufgabenstellungen unterschiedlich gut (vgl. Abbildung 3.2.3.4./3). Strebt man eine Strukturanalyse an, steht als weitere Entscheidung an, ob alle Objekte klassifiziert werden sollen oder ob - zugunsten von eindeutigen Klassengrenzen - unklassifizierte Objekte bleiben (nicht-exhaustive Clusteranalyse).

- Methodenverknüpfung als Ziel der Analyse?

Zum Teil wird die Clusteranalyse nur als Vorstudie für andere Verfahren verwendet. In diesem Fall empfiehlt es sich, spezielle Algorithmen zu verwenden; beispielsweise existieren Verfahren, die eine klassenspezifische Regressionsanalyse unterstützen¹⁾. Auch die erforderliche Genauigkeit kann je nach geplanter Methodenverknüpfung unterschiedlich sein; werden im folgenden nur noch die Schwerpunkte benötigt, ist eine geringere Genauigkeit erforderlich als wenn die Gruppenzugehörigkeit jedes Objekts entscheidend ist.

- Überlappungen gewünscht?

Für manche Anwendungen, beispielsweise die Clusteranalyse von Dokumenten für ein Information-Retrieval-System, ist es sinnvoll, ein Objekt in mehrere Cluster einzugruppieren; der Aufwand der Analyse erhöht sich entsprechend.

¹⁾ Vgl. z. B. Bock, H., Automatische Klassifikation, a.a.O., S. 195 ff.

- Sind Nebenbedingungen zu beachten?

Weitgehend ungelöst ist das Problem der Nebenbedingungen bei einer Clusteranalyse; der Anwender möchte aber gerade hier seine Vorstellungen einbringen. Bei der Suche von optimalen Standorten ist es erforderlich, zusammenhängende Cluster zu erhalten; für andere Aufgabenstellungen dürfen die Gruppen nicht zu groß oder zu klein sein. Eine bestimmte Mindesthomogenität innerhalb der Gruppen ist zu fordern, wenn der Schluß vom Cluster auf alle Objekte mit hoher Sicherheit gezogen werden muß.

Während die Zielvorstellungen des Analytikers im wesentlichen bestimmen, welches Verfahren sinnvoll ist, schränken die Eigenschaften der verwendeten Daten häufig die Zulässigkeit einzelner Methoden ein; im einzelnen ist zu beachten:

- Welcher Art sind die Daten?

Je nachdem, ob für jedes Objekt alle Merkmalsausprägungen gegeben sind oder ob Ähnlichkeitsbeziehungen zwischen den Objekten vorliegen, sind unterschiedliche Verfahren einsetzbar. Dabei ist insbesondere zu beachten, daß Clusteralgorithmen in der Regel auf Distanzmaße ausgerichtet und Ähnlichkeiten noch in Distanzen zu transformieren sind.

- Zahl der Objekte und der Merkmale?

Die Zahl der Objekte und der Merkmale schließt häufig bestimmte Verfahren von vornherein aus. So ist teilweise eine große Zahl von Objekten notwendig, um ein Ergebnis mit ausreichender Sicherheit zu liefern; andere Verfahren erreichen bei einer Vielzahl von Objekten solche Laufzeiten, daß ihr Einsatz unwirtschaft-

lich ist. Auch die Auswahl zwischen den Methodengruppen hängt von diesen Zahlen ab: Bei vielen Objekten sind partitionierende Verfahren tendenziell vorzuziehen, bei vielen Merkmalen hierarchische Verfahren (vgl. Abschnitt 3.2.3.1.).

- Skalenniveau der Merkmale?

Die Skalierung der Merkmale ist von besonderer Bedeutung; sie fand deshalb auch bereits in ANCA starkes Gewicht (vgl. Abschnitt 3.2.3.). Sehr wichtig ist auch die Vergleichbarkeit der Skalen; ist sie nicht gegeben, kann eine Standardisierung als Baustein sinnvoll sein.

- Bedeutung einzelner Merkmale?

Dieses Kriterium ist ausschlaggebend dafür, ob die Merkmale gewichtet oder ungewichtet in die Analyse eingehen; zum Teil wird damit auch eine automatische Vorselektion von Verfahren durchgeführt, weil nicht alle Algorithmen eine Gewichtung zulassen.

Externe Faktoren sind ausschließlich durch den Anwender bedingt; sie gelten in ähnlicher Form auch für andere Methoden als die Clusteranalyse:

- Erlaubter Aufwand

Anwender sind häufig nicht bereit, sich intensiv mit bestimmten Verfahren auseinanderzusetzen; bei Initialisierung der Methode und Darstellung der Ergebnisse dürfen deshalb nicht zu viele Entscheidungen offen bleiben, die Wissen und Erfahrung des Analytikers ansprechen.

- Budget-Restriktion

Häufig ist vorgegeben, wieviel eine Analyse kosten darf; bestimmte Verfahren verbieten sich bei niedri-

gem Budget von selbst, bei anderen sind Nutzenabwägungen sinnvoll.

- Akzeptanzvorstellung des Anwenders

Bei einer stark erweiterten Methodensammlung kann nicht mehr davon ausgegangen werden, daß alle Methoden in etwa gleich verständlich sind. In den Auswahlprozeß müssen deshalb das Methodenwissen des Analytikers und seine Bereitschaft zur Anwendung von Verfahren, die er nicht vollständig versteht, eingehen (vgl. dazu Abschnitt 3.2.2.).

5.2.2.3.2. Gestaltung des Mustererkennungsprozesses

Bei vergleichsweise unübersichtlichen Merkmalskombinationen und einer großen kombinatorischen Vielfalt möglicher Maßnahmen bietet sich häufig ein Mustererkennungsprozeß¹⁾ an, der eine Verbindung von den beobachteten Merkmalen ("Eingangsmuster") zu einer geeigneten Handlungsempfehlung ("Ausgangsmuster") herstellt²⁾. In unserem Fall gilt es, die Kriterien eines Clusteranalyseproblems zu beobachten und dafür (im Idealfall eine) Methode(n) zur Problemlösung vorzuschlagen.

Bei dieser Anwendung liegen einfache abstrakte Muster³⁾ vor; zu ihrer Verarbeitung gehören üblicherweise die Phasen Vorverarbeitung, Merkmalsextraktion und Klassi-

¹⁾ Zu einer Einführung in die Mustererkennung vgl. Niemann, H., Mustererkennung - Einführung und Überblick, Informatik Spektrum 2 (1979), S. 12. ff.; eine umfassende Übersicht gibt Niemann, H., Methoden der Mustererkennung, Frankfurt 1974.

²⁾ Vgl. dazu Mertens, P., Die Theorie der Mustererkennung in den Wirtschaftswissenschaften, Zeitschrift für betriebswirtschaftliche Forschung 29 (1977) 12, S. 777 ff.

³⁾ Zur Abgrenzung vgl. ebenda.

fizierung, zur Bestimmung der Klassencharakteristika kommt eine Lernphase hinzu. In unserem Fall entfällt die Phase der Vorverarbeitung, weil alle Merkmale in einem interaktiven Prozeß direkt erfaßt werden.

Als Merkmale finden die im vorigen Abschnitt definierten anwenderorientierten Kriterien Verwendung. Einige davon können unmittelbar übernommen, andere müssen noch stärker operationalisiert werden. Beispielsweise ist es bei überlappender Clusteranalyse erforderlich, den Grad der Überlappung als Merkmal zu definieren; die Nebenbedingung "Mindesthomogenität" muß so beschrieben werden, daß sich eine Ähnlichkeitsschranke oder ein Abbruchkriterium ableiten läßt¹⁾.

Die Klassifizierungsphase ordnet aufgrund der gespeicherten Klassencharakteristika das Muster einer Klasse zu. In diesem Schritt muß auch Übereinstimmung zwischen dem Anforderungsprofil des Anwenders, beschrieben mit den oben definierten Kriterien, und den statistischen Eigenschaften der einzelnen Verfahren hergestellt werden, ohne daß diese nach außen in Erscheinung treten müssen. So sind beispielsweise bei der Suche nach Außenseitern Verfahren auszuschließen, die gleiche Varianzen innerhalb der Klassen voraussetzen. Ein weiterer Anwendungsfall ist die Clusteranalyse als Vorstufe zu klassenspezifischer Regressionsanalyse; hier muß ein Verfahren gesucht werden, das die gefundenen Gruppen durch Hyperebenen charakterisiert.

Dieses Vorgehen setzt voraus, daß die Clusteranalyseverfahren selbst klassifiziert sind und daß zwischen den Gruppen Trennlinien gefunden werden. Dabei soll

¹⁾ Zur Problematik dieser Forderung vgl. die Fußnote 2 auf Seite 35.

nicht über den heuristischen Charakter der meisten Verfahren hinweggesehen werden. Die Entscheidung zwischen der Notwendigkeit zu selektieren und der Gefahr, Anwendungsmöglichkeiten zu stark einzuschränken, wird häufig nach Ermessen gefällt werden müssen. Gegenüber der zur Zeit teilweise beobachteten Auswahl nach zufälliger Verfügbarkeit stellt dies trotzdem eine Verbesserung dar.

Der einfachste Fall der Klassifikation ist der Schablonenvergleich, der alle Merkmale der Problemstruktur und die Klassencharakteristika auf Identität prüft. Mit diesem Vorgehen ist allerdings eine unpraktikable Zahl von Klassen verbunden; es ist deshalb notwendig, in einer Lernphase geeignete Klassencharakteristika zu definieren. Dazu bietet sich ein Verfahren des überwachten Lernens an, das Erfahrungen aus Verfahrensvergleichen und von praktischen Anwendungen einbeziehen kann. Beispielsweise ist es bei der erwähnten Ähnlichkeitsschranke möglich, versuchsweise einen Schwellwert zu definieren. Nach jeder Anwendung wird in das System eingegeben, ob sie erfolgreich war oder nicht; gegebenenfalls wird daraufhin die Schranke verändert.

Das Mustererkennungssystem selbst kann hierarchisch oder parallel gestaltet sein. Da es nicht sinnvoll ist, für eine Anwendung alle möglichen - auch nicht relevanten - Parameter zu erfragen, ist eine hierarchische Vorgehensweise vorzuziehen. Beispielsweise kann das System in der ersten Phase die Akzeptanzbereitschaft des Benutzers prüfen; ist dieser nicht bereit, ihm nicht einschichtige Verfahren anzuwenden, entfällt eine Reihe von Methoden und damit von Fragen an den Benutzer, die sich auf diese nicht angebotenen Verfahren beziehen.

Eine Präzisierung dieses Ansatzes ist noch verfrüht. In der nahen Zukunft müssen die Kriterien zur Methodenwahl und die Kategorisierung der Verfahren im Vordergrund stehen. Die Bereitschaft, auch negative Erfahrungen mit bestimmten Methoden in praktischen Anwendungsfällen weiterzugeben, und umfangreiche gezielt aufgebaute Vergleichsstudien können hierzu viel beitragen.

5.2.3. Ergebnisbehandlung

Die Routinen zur Darstellung und Interpretation der Ergebnisse reichen aus, um die meisten Anforderungen erfüllen zu können; allenfalls könnte der in Abschnitt 5.1.3. angesprochene verstärkte Einsatz graphischer Terminals empfohlen werden.

Für Bewertung und Vergleich von Clusterergebnissen scheinen Verbesserungen möglich, obwohl das in ANCA gewählte pragmatische Vorgehen anerkannt ist¹⁾. Zum Vergleich von Verfahren wurde eine Vielzahl von Maßen geschaffen²⁾; da sich aber keines von ihnen durchsetzen konnte, sollte vorerst auf eine Implementierung verzichtet werden. Sinnvoll wäre auch ein Test auf die "richtige" Klassenanzahl mit Hilfe der Originaldaten³⁾; allerdings sind auch hier keine problemlos anwendbaren Tests bekannt.

¹⁾ Vgl. Dubes, R. und Jain, A. K., Models ..., a.a.O., S. 153: "A user of cluster algorithms interested in validity would be well advised to apply several clustering approaches and check for common clusters instead of searching for a technical measure of validity for an individual clustering."

²⁾ Vgl. z. B. Rand, W. M., a.a.O.; Lessig, V. P., Comparing Cluster Analysis with Cophenetic Correlation, Journal of Marketing Research 9 (1972), S. 82 ff.; Fangmeyer, H. und Gloden, R., Bewertung und Vergleich von Klassifikationsergebnissen bei automatischen Verfahren, in: Gesellschaft für Klassifikation (Hrsg.), Kooperation in der Klassifikation, Studien zur Klassifikation, Band 2, Frankfurt 1978, S. 147 ff.

³⁾ Vgl. Bock, H. H., Automatische Klassifikation, a.a.O., S. 193 f.

5.3. Diskussion des DV-Konzepts

5.3.1. Offenheit des Systems

Voraussetzung für die angesprochenen Erweiterungsmöglichkeiten von ANCA ist Offenheit des Systems in bezug auf Methoden und Daten.

Wie in 4.2.2. angesprochen, wurde bei der Gestaltung das Prinzip der Modularität berücksichtigt. Dennoch sind größere Änderungen bei einem Ausbau der Methodensammlung von ANCA nicht zu vermeiden. Im einzelnen werden erforderlich:

- Erstellen eines spezifischen Unterprogramms
- Einfügen des Unterprogrammaufrufs mit allen Parametern
- Veränderung der Auswahlmenüs zur Methodenwahl
- Änderung der Entscheidungsregeln zur Methodenwahl.

Wünschenswert wäre eine Erweiterung des Methodenvorrats, ohne jedesmal die Programme ändern, neu übersetzen und montieren zu müssen. Eine Lösung ist das dynamische Zuladen von Unterprogrammen zur Laufzeit¹⁾. Die Adreßbezüge zwischen Haupt- und Unterprogramm werden erst beim Aufruf des Unterprogramms erstellt und nicht beim Laden des Hauptprogramms. Leider kennt die zur Verfügung stehende Rechananlage TR440 diese Möglichkeit nicht.

Die Verarbeitung neuer Daten ist durch die generelle Datenschnittstelle gewährleistet. Allerdings können keine

¹⁾ Das dynamische Zuladen stellt für Schips eine unverzichtbare Bedingung für ein Methodenbankkonzept dar; vgl. Schips, B., Ein Beitrag ..., a.a.O.

Strukturen bearbeitet werden, die von der Blockstruktur (ein Satz je Objekt) abweichen. Eventuell sollte die Eingabe einer Ähnlichkeitsmatrix möglich sein; diese könnte mit anderen Methoden, beispielsweise mit der Mehrdimensionalen Skalierung¹⁾, erstellt werden.

5.3.2. Dialogablauf

Wie in Abschnitt 4.3.2. dargestellt, wird momentan nicht zwischen Laien- und Expertenmodus im Dialogablauf unterschieden. Eine Erweiterung des Methodenvorrats hätte möglicherweise zur Folge, daß erklärende Texte aus den Dialogbildern herausgenommen werden müßten; die Einführung verschiedener Modi sollte dann erneut geprüft werden²⁾. Dabei sollte keine starre Unterscheidung gewählt werden³⁾; neben der Möglichkeit zum Moduswechsel an jeder Stelle⁴⁾ ist vor allem an ein automatisches Umschalten auf den ausführlichen Modus zu denken, wenn der Benutzer sehr lange zögert⁵⁾.

ANCA ist bisher darauf ausgerichtet, eine Clusteranalyse "in einem Guß" durchzuführen; erst wenn ein Ergebnis vorliegt und abgespeichert ist, kann zu einem späteren Zeit-

-
- 1) Zu dieser Verbindung vgl. z. B. Dichtl, E. und Schobert, R., Mehrdimensionale Skalierung - Methodische Grundlagen und betriebswirtschaftliche Anwendungen, München 1978, S. 64.
- 2) Vgl. allerdings die Beobachtung von Cabus, daß sich die Erstellung einer ausführlichen Kommunikationsform nicht lohnt: Cabus, E., a.a.O., S. 271.
- 3) Vgl. z. B. die Kritik von Szyperski, N., Realisierung von Informationssystemen in deutschen Unternehmungen, in: Müller-Merbach, H. (Hrsg.), Quantitative ..., a.a.O., S. 77.
- 4) Vgl. z. B. Bodendorf, F., Renninger, W. und Schaller, K., Zwei neue ..., a.a.O., S. 48.
- 5) Vgl. z. B. den Hinweis von Brockhoff auf das System IDA: Brockhoff, K., Prognoseverfahren für die Unternehmensplanung, Wiesbaden 1977, S. 156.

punkt die Analyse fortgesetzt werden. Es ist zu prüfen, ob die bei CLUSTAN vorgesehene RESTART-Prozedur¹⁾, die an jeder Stelle ein Unterbrechen des Dialogflusses ermöglicht, auch bei ANCA eingeführt werden sollte.

Sinnvoll erscheint es auch, die Möglichkeit des Schnellgangs durch Einrichtung eines generellen Kommandospeichers zu erweitern. Dies setzt ein umfassendes Menu voraus, in das alle Eingaben des gesamten Dialogs eingetragen werden können. Das System arbeitet diese Kommandos ab, solange sie angegeben und formell richtig sind; einzelne fehlende Parameter könnte es intelligent nachfragen²⁾.

Eine weitergehende Variante der Anwenderunterstützung im Dialogablauf ist eine benutzereigene Sprache³⁾ für Dialogbilder und Eingaben. In Anbetracht der unterschiedlichen Anwendergruppen der Clusteranalyse, die zum Teil mit Begriffen der Graphentheorie, der Statistik, der Mustererkennung oder aus Anwendungsgebieten - z. B. der Betriebswirtschaftslehre - arbeiten, könnte sich dies für ein System zur Clusteranalyse anbieten. Da ANCA für Anwender aus dem wirtschaftswissenschaftlichen Bereich konzipiert ist, wurde diese Anregung nicht weiter verfolgt.

Die Einführung solcher Verbesserungen setzt eine ausführliche Analyse des Dialogablaufs voraus. Dabei können die vorhandenen Dialogprotokolle sinnvolle Dienste leisten. Es lassen sich z. B. ableiten:

¹⁾ Vgl. Schäffer, J. B., a.a.O.; allerdings wäre es bei ANCA nicht möglich, wie in CLUSTAN dem Anwender selbst die Verantwortung für eine Sicherung der entsprechenden Daten zu überlassen.

²⁾ Vgl. z. B. die Vorgehensweise bei IPROS: Diepenbrock, F., Einführung in das interaktive Prognose-System IPROS, Teil 1, München 1975, S. 12.

³⁾ Vgl. Kogon, R. u. a., User Specialty Languages, Technical Report 75.12.007, Heidelberg 1975.

- fehlerhafte Eingaben in einzelnen Phasen des Dialogs
- unbefriedigte HILFE-Rufe
- typische Dialogabläufe (vgl. Abschnitt 5.3.3.)
- zeitliches Verhalten im Dialog¹⁾.

5.3.3. Einbau von Benutzermodellen

Bei der Methodenanwendung lassen sich häufig bestimmte Verhaltensmuster finden, die in verschiedenen Analysen immer wiederkehren. Diese können im einfachsten Fall der typischen Vorgehensweise eines Anwenders oder einer Anwendergruppe entsprechen. Denkt man diesen Ansatz weiter, können sich dahinter aber auch Richtlinien verbergen, "welche Abfolge von Verfahren für die Bearbeitung bestimmter ... immer wiederkehrender Probleme (z. B. die Vorbereitung von Segmentierungsentscheidungen) zu empfehlen sind"²⁾. Der Einstieg in eine Methodenbank sollte auch über solche stärker anwendungsbezogene Fragestellungen möglich sein³⁾.

Die Abspeicherung solcher "Verfahrenshierarchien", die teilweise mehrere komplexe Verfahren umfassen können, ist in einer speziellen Modelldatei möglich⁴⁾. Sie werden mit einer Bezeichnung versehen, mit der sie auch auf-

1) Dazu ist es erforderlich, bei jedem Dialogschritt die Uhrzeit mit auszugeben. Dies ist nicht realisiert, mit geringem Aufwand aber einzufügen.

2) Topritzhofer, E., Marketing - Neue Ergebnisse aus Forschung und Praxis, Wiesbaden 1978, S. 20.

2) Vgl. ebenda. Topritzhofer erwartet von einer benutzerfreundlichen Methodenbank genau diese Leistung. Allerdings räumt er selbst ein, daß die Erstellung solcher Methodenbanken der Mitwirkung von Marketingspezialisten bedarf, die sich aber erst seit kurzer Zeit mit diesen Problemen befassen.

4) Vgl. z. B. die bereits bestehende Lösung bei METHAPLAN: Rickert, R., a.a.O., S. 243.

gerufen werden. Bei der Bearbeitung wird in der Regel auf Eingaben verzichtet; an bestimmten Stellen können Wahlmöglichkeiten vorgesehen sein, ansonsten läuft die Verarbeitung bis zum Beginn der Interpretationsroutinen automatisch ab. Voraussetzung für eine Realisierung dieses Ansatzes ist die Implementierung des im vorigen Abschnitt angesprochenen generellen Kommandospeichers über das gesamte System MADAS hinweg.

5.4. Diskussion des Umfeldes von ANCA

5.4.1. Verbund von Systemen zur Clusteranalyse

Zunehmender Komfort in einer Methodenbank muß mit Verzicht auf Allgemeingültigkeit erkaufte werden; benutzerfreundliche Systeme lösen deshalb gemäß dem 80/20-Prinzip nur etwa 80 % der anfallenden Probleme. Gerade bei dem Variantenreichtum der Clusteranalyse wird es immer erforderlich sein, auf mehrere Methodenbanksysteme zurückzugreifen. In dieser Situation gibt es verschiedene Möglichkeiten, wie dem "emanzipierten" Benutzer in ANCA begegnet werden kann:

1. Er erhält keine Unterstützung oder nur den allgemeinen Hinweis auf Pakete wie CLUSTAN.
2. Er erhält einen gezielten Hinweis auf eine Prozedur, z. B. aus CLUSTAN. Voraussetzung für diese Variante ist eine Ausdehnung des Methodenauswahlprozesses auf mehrere Pakete¹⁾.

¹⁾ Diese methodenbankübergreifende Methodenauswahl wird derzeit in einem von der Informatik-Forschungsgruppe VIII der Universität Erlangen-Nürnberg gemeinsam mit dem Regionalen Rechenzentrum Erlangen durchgeführten und von der Deutschen Forschungsgemeinschaft geförderten Projekt näher untersucht.

3. Er erhält - wie bei Variante 2 - einen Hinweis und, falls er die Methode auswählt, ein ablauffähiges Steuerkartendeck für das betreffende System¹⁾.

Eine Erweiterung von ANCA um eine der zwei letztgenannten Alternativen ist momentan nicht geplant; dies sollte erst nach einer erfolgreichen Akzeptanzbetrachtung des bestehenden Systems ins Auge gefaßt werden. Darüber hinaus ist zu prüfen, ob die Ergebnisse, die mit einem anderen System erstellt wurden, mit den Interpretationsroutinen von ANCA analysiert werden können; der Aufwand hierfür scheint allerdings beträchtlich.

5.4.2. Ergänzung um Hilfsroutinen

Die Einbettung von ANCA in das System MADAS erfolgte, um Datenanalysen mit größtmöglichem Komfort durchführen zu können. Neben der eigentlichen Methodenverarbeitung gehören dazu vor allem die Bausteine zur Aufbereitung, Auswahl und Veränderung der Daten.

Alle genannten Teilaufgaben werden in der ersten hier beschriebenen Version bei der fest angeschlossenen Panneldatenbank unterstützt. Für Forschungsarbeiten sind dagegen Erweiterungen denkbar; insbesondere ist zu berücksichtigen, daß der Analytiker in der Regel selbst für die Erfassung seiner Daten und Beschreibungsinformationen zu sorgen hat.

¹⁾ Zu dieser Vorgehensweise vgl. das System SAMBA, das Steuerkarten für SPSS erzeugt: Bodendorf, F., Realisierung ausgewählter Entwurfskriterien in der statistischen Methodenbank SAMBA, Vortrag bei der Tagung "Computer am Arbeitsplatz", Linz 1979 (unveröffentl. Manuskript).

Die erforderlichen Merkmalsstammsätze, Steuerparameter und Dialogbilder zur Verarbeitung einer beliebigen Datei können in einem Dialogprogramm definiert werden, das erstellt und erprobt wurde¹⁾. Dagegen wird die Erfassung der Daten selbst momentan nicht unterstützt. Hier bietet sich die Verwendung eines Generators an, der in der Forschungsgruppe programmiert wurde²⁾. Mit geringem Aufwand läßt sich eine individuelle Erfassungsroutine erstellen, die neben der Bereitstellung der Daten selbst die Beachtung des vorgeschriebenen Formats gewährleistet. Auch die Definition von Stichproben nach logischen Kriterien wird momentan nur für die Pilotdatenbank von MADAS unterstützt³⁾; hier wäre ein allgemeiner Baustein sinnvoll.

¹⁾ Vgl. Pansch, P., Erstellung und Test von Programmen zur Integration der Methodenbanksysteme MADAS und ANCA, Studienarbeit, Erlangen 1979.

²⁾ Das Programm wurde von F. Bodendorf erstellt; es ist nur intern im Einsatz.

³⁾ Vgl. Scholz, P., Konzeption und Realisierung eines Datenbankdialogsystems zur mathematisch-statistischen Auswertung von Marktforschungsdaten, Diplomarbeit, Erlangen 1975.

6. Ausblick

Die vorliegende Arbeit zeigt, daß auch anspruchsvolle Verfahren in Methodenbanken integriert und angewandt werden können. Krautter wagt die Prognose, "..., daß eine bessere Ausschöpfung des Potentials computergestützter Methoden der Entscheidungsfindung - verbunden mit unternehmerischer Intuition - eine wichtige Voraussetzung für ein erfolgreicher Marketing in den 80er Jahren sein wird."¹⁾

Methodenbanken werden sich voraussichtlich noch stärker durchsetzen. Damit gewinnt der in Abschnitt 1.1. angesprochene Aspekt der wirtschaftlichen Realisierung von Methodenbanken an Bedeutung. Einen möglichen Weg zur Lösung dieser Problematik will das System KARAMBA²⁾ aufzeigen. Es unterscheidet zwischen vier Nutzergruppen, angefangen bei denjenigen, die Methoden erstellen und pflegen, bis hin zu Anwendern, die sich einer fertigen Methodenbank bedienen. Auf jeder Ebene werden die vorhandenen Bausteine so aufbereitet, daß sie zum einen unverändert in der nächsten Stufe eingesetzt werden können, daß zum anderen aber unterschiedliche Systeme gestaltet werden können. Das Ziel der Offenheit wird somit gut realisiert.

Werden bei der Systemerstellung die gesammelten Kriterien zur Gestaltung anwenderbezogener Methodenbanken verwendet, steht einem Durchbruch von Methodenbanken nichts mehr im Wege.

¹⁾ Krautter, J., Marketing-Modelle - Stagnation ohne Ende?, Absatzwirtschaft 22 (1979) 9, S. 91 ff., hier S. 97.

²⁾ Hüber, R., Dittrich, K. R. und Lockemann, P. C., Das KARAMBA Methodenbanksystem, in: Böhling, K. H. und Spies, P. P. (Hrsg.), Proceedings der 9. Jahrestagung der Gesellschaft für Informatik, Berlin-Heidelberg-New York 1979, S. 322 ff.

Literaturverzeichnis

Allerbeck, K. R., Datenanalyse, in: Koolwijk, J. van und Wieken-Mayser, M. (Hrsg.), Techniken der empirischen Sozialforschung, München 1974.

Anderberg, M. R., Cluster Analysis for Applications, New York 1973.

Bock, H. H., Automatische Klassifikation, Göttingen 1974.

Bock, H. H., Programme zur Clusteranalyse, Statistical Software Newsletter 2 (1976), S. 72 ff.

Bodendorf, F., Realisierung ausgewählter Entwurfskriterien in der statistischen Methodenbank SAMBA, Vortrag bei der Tagung "Computer am Arbeitsplatz", Linz 1979 (unveröffentl. Manuskript).

Bodendorf, F., Renninger, W. und Schaller, K., Zwei neue Methodenbanksysteme für statistische Anwendungen, Arbeitsbericht des Instituts für Mathematische Maschinen und Datenverarbeitung der Universität Erlangen-Nürnberg 11 (1978) 22.

Böcker, F., Zur Akzeptanz formaler Entscheidungsmodelle im Marketing, Wirtschaftswissenschaftliches Studium 6 (1977), S. 305 ff.

Böcker, F., Modellbezogene Akzeptanzprobleme formaler Entscheidungsmodelle im Marketing, in: Müller-Merbach, H. (Hrsg.), Quantitative Ansätze in der Betriebswirtschaftslehre, München 1978, S. 227 ff.

Böhler, H., Der Beitrag von Konsumententypologien zur Marktsegmentierung, Die Betriebswirtschaft 37 (1977), S. 447 ff.

Böhling, K. H. und Spies, P. P. (Hrsg.), Proceedings der 9. Jahrestagung der Gesellschaft für Informatik, Berlin-Heidelberg-New York 1979.

Bradl, F., Test und Erweiterung des Systems zur Clusteranalyse im Rahmen der interaktiven Methodenbank MADAS, Diplomarbeit, Nürnberg 1978.

Braunschweig, E. (Hrsg.), Typologien, o. O. 1975.

Breitung, A., Marketing-Informationssysteme, Meisenheim 1975.

Brockhoff, K., Prognoseverfahren für die Unternehmensplanung, Wiesbaden 1977.

Brockhoff, K., Dinkelbach, W., Kall, P., Pressmar, D. B. und Spicher, K. (Hrsg.), Proceedings in Operations Research 7, Würzburg 1978.

Buttler, G., Methoden zur Abgrenzung regionaler Arbeitsmärkte, Köln 1974.

Cabus, E., Ansätze für ein Marketing-Informationssystem, in: Haedrich, G. (Hrsg.), Operationale Entscheidungshilfen für die Marketingplanung, Berlin-New York 1977, S. 269 ff.

Cormack, R. M., A Review of Classification, Journal Royal Stat. Soc., Series A, 134 (1971), S. 321 ff.

Corsten, L. C. A. und Hermans, J. (Hrsg.), COMPSTAT 1978: Proceedings in Computational Statistics, 3. Symposium held in Leiden, Wien 1978.

Cunningham, K. M. und Ogilvie, J. C., Evaluation of hierarchical Grouping Techniques: a preliminary Study, Computer Journal 15 (1972), S. 209 ff.

Dathe, H. N., Mertens, P., Peschanel, F. D., Späth, H. und Zimmermann, H. J. (Hrsg.), Proceedings in Operations Research 6 (1976).

Degens, P. und Federkiel, H., A Monte Carlo Study on Agglomerative Large Sample Clustering, in: Corsten, L. C. A. und Hermans, J. (Hrsg.), COMPSTAT 1978: Proceedings in Computational Statistics, 3. Symposium held in Leiden, Wien 1978, S. 246 ff.

Dichtl, E. und Schobert, R., Mehrdimensionale Skalierung - Methodische Grundlagen und betriebswirtschaftliche Anwendungen, München 1978.

Dickhoven, S. (Hrsg.), Modellierungssoftware, St. Augustin 1976.

Diepenbrock, F., Einführung in das interaktive Prognose-System IPROS, Teil 1, München 1975.

Diller, H., Marketing-Informationssysteme, Wirtschaftswissenschaftliches Studium 5 (1976), S. 97 ff.

Dixon, W. J. (Hrsg.), BMD - Biomedical Computer Programs, Berkeley-Los Angeles 1970.

Dixon, W. J. (Hrsg.), Biomedical Computer Programs - P-Series, Berkeley-Los Angeles 1977.

Dubes, R. und Jain, A. K., Clustering Techniques: The User's Dilemma, Pattern Recognition 8 (1976), S. 247 ff.

Dubes, R. und Jain, A. K., Models and Methods in Cluster Validity, in: IEEE (Hrsg.), Proceedings der Conference on Pattern Recognition and Image Processing, Chicago 1978.

Eckardt, Th., Eine Prognosemethodenbank für Kleinrechner - Konzept, Stand der Entwicklung und erste Erfahrungen, Vortrag anlässlich der 1. Sitzung der Arbeitsgruppe "Prognoseverfahren" der Deutschen Gesellschaft für Operations Research, Berlin 1978 (unveröffentl. Manuskript).

Erbe, R. und Walch, G., An Interactive Guidance System for Method Libraries, IBM Publikation 75.004.001, Heidelberg 1975.

Esprester, A. C., Die Entwicklung einer Methodenbank und einer Methodenbanksprache, Angewandte Informatik 20 (1978), S. 203 ff.

Everitt, B., Cluster Analysis, London u. a. 1974.

Faber, E. und Nollau, W., Über ein Verfahren zur automatischen Klassifikation, Darmstadt 1969.

Fangmeyer, H. und Gloden, R., Bewertung und Vergleich von Klassifikationsergebnissen bei automatischen Verfahren, in: Gesellschaft für Klassifikation (Hrsg.), Kooperation in der Klassifikation, Band 2, Frankfurt 1978, S. 147 ff.

Farley, J. U. und Howard, J. A. (Hrsg.), Control of "Error" in Market Research Data, Lexington-Toronto-London 1975, S. 109 ff.

Forst, H., Zur Klassifizierung von Städten nach wirtschafts- und sozialstatistischen Strukturmerkmalen, Würzburg 1974.

Forst, H. T. und Vogel, F., Y-Group: Ein Fortran-IV-Programm zur hierarchischen Klassifizierung von Merkmalsträgern und Merkmalen, Köln 1972 (unveröffentl. Manuskript).

Forst, H. T. und Vogel, F., Diagnose von Klassifikationsergebnissen, Köln-Kiel 1978 (unveröffentl. Manuskript).

Frank, R. E. und Green, P. E., Numerical Taxonomy in Marketing Research, Journal of Marketing Research 5 (1968), S. 83 ff.

Geschka, H., Implementierungsprobleme bei der Anwendung von Ideenfindungsmethoden in der Praxis der Unternehmen, in: Pfohl, H.-C. und Rürup, B. (Hrsg.), Anwendungsprobleme moderner Planungs- und Entscheidungstechniken, Königstein 1978, S. 159 ff.

Gesellschaft für Klassifikation (Hrsg.), Kooperation in der Klassifikation, Studien zur Klassifikation, Band 2, Frankfurt 1978.

Gesellschaft für Klassifikation (Hrsg.), Klassifikation und Erkenntnis III, Studien zur Klassifikation, Band 6, Frankfurt 1979.

Graham, R. J., Is Management Science arcane?, Interfaces 7 (1977), S. 63 ff.

Green, P. E. und Rao, V. R., A Note on Proximity Measures and Cluster Analysis, Journal of Marketing Research 6 (1969), S. 359 ff.

Greenberg, B. A., Goldstucker, J. L., Bellinger, D. N., What Techniques are used by Marketing Researchers in Business?, Journal of Marketing 41 (1977), S. 62 ff.

Haedrich, G. (Hrsg.), Operationale Entscheidungshilfen für die Marketingplanung, Berlin-New York 1977.

Hamacher, B. und Preiser, K., Eine Infrastrukturtypologie am Beispiel des Landes Bremen, in: Späth, H. (Hrsg.), Fallstudien zur Clusteranalyse, München 1977, S. 147 ff.

Hansen, H. R. (Hrsg.), Computergestützte Marketingplanung, München 1974.

Hartigan, J. A., Clustering algorithms, New York 1975.

Heinzelbecker, K., Partielle Marketing-Informationssysteme, Moduln für computergestützte Entscheidungen in Marketing, Verkauf, Vertrieb, Zürich-Frankfurt 1977.

Heinzen, H. J. und Konkart, R., GRADAS - ein interaktives Informationssystem auf der Grundlage einer relationalen Datenbank, Interner Bericht der AEG, o. O., o. J.

Herrmann, U., Analyse eines Analytischen Arbeitsplatzbewertungs- und Entlohnungssystems, Diplomarbeit, Nürnberg (in Arbeit).

Hüber, R., Dittrich, K. R. und Lockemann, P. C., Das KARAMBA Methodenbanksystem, in: Böhling, K. H. und Spies, P. P. (Hrsg.), Proceedings der 9. Jahrestagung der Gesellschaft für Informatik, Berlin-Heidelberg-New York 1979, S. 322 ff.

Hüber, R. und Lockemann, P. C., Information Protection by Method Base Systems, in: Yao, S. B. (Hrsg.), Proceedings der Fourth International Conference on Very Large Data Bases, Berlin 1978, S. 420 ff.

IBM (Hrsg.), SSP III - Programmer's Manual, IBM-Form 6H 20-0205-4.

ICPR (Hrsg.), OSIRIS III Manual, Ann Arbor 1974.

IEEE (Hrsg.), Proceedings der Conference on Pattern Recognition and Image Processing, Chicago 1978.

Kamen, J., Quick Clustering, Journal of Marketing Research 7 (1970), S. 199 ff.

Köhler, R. und Uebele, H., Planung und Entscheidung im Absatzbereich industrieller Großunternehmen, Aachen 1977.

Kogon, R. u. a., User Specialty Languages, Technical Report 75.12.007, Heidelberg 1975.

Kohlas, I., Seifert, O., Stähly, P. und Zimmermann, H. J. (Hrsg.), Proceedings in Operations Research 5, Würzburg 1975.

Koolwijk, J. van und Wieken-Mayser, M. (Hrsg.), Techniken der empirischen Sozialforschung, München 1974.

Krautter, J., Marketing-Modelle - Stagnation ohne Ende?, Absatzwirtschaft 22 (1979) 9, S. 91 ff.

Krautter, J. und Scholz, P., Entwicklung computergestützter Marketing-Informationssysteme, Die Betriebswirtschaft 38 (1978), S. 231 ff.

Lades, R. und Schläger, W., CLAN, ein EDV-Programm zur Clusteranalyse für die Rechenanlage CD 3300, Arbeitspapiere des Betriebswirtschaftlichen Instituts der Friedrich Alexander-Universität Erlangen-Nürnberg, Heft 26, Nürnberg 1974.

Landgrebe, K. P., Konstruktionsprinzipien und Erkenntnisgehalte neuerer Bevölkerungsmodelle für das Marketing, in: Braunschweig, E. (Hrsg.), Typologien, o. O. 1975, S. 113 ff.

Leece, J. und Parker, F., Use and Misuse of SPSS, Software - Practice and Experience 8 (1978), S. 301 ff.

Leisch, E., Bestandsaufnahme zum Einsatz multivariater Verfahren bei wirtschaftlichen Entscheidungsprozessen, Diplomarbeit, Nürnberg 1977.

Lessig, V. P., Comparing Cluster Analysis with Cophenetic Correlation, Journal of Marketing Research 9 (1972), S. 82 ff.

Little, J. D. C., Models and Managers: The Concept of a Decision Calculus, Management Science 16 (1970), S. B-466 ff.

Lorenz, N., Ein Programmpaket zur Cluster-Analyse im Rahmen des Dialogsystems MADAS unter besonderer Berücksichtigung der Unterstützung des Benutzers, Diplomarbeit, Nürnberg 1977.

Makridakis, S., Hodgson, A. und Wheelwright, S. C., An Interactive Forecasting System, Journal of the HP Educational Users Group 1 (1973) 2, S. 7 ff.

Mertens, P., Möglichkeiten und Grenzen eines branchenorientierten Marktdatenbanksystems (MADAS), in: Hansen, H. R. (Hrsg.), Computergestützte Marketingplanung, München 1974, S. 181 ff.

Mertens, P., Die technische Gestaltung von Bildschirm-Dialogen, Online 15 (1977), S. 670 ff.

Mertens, P., Die Theorie der Mustererkennung in den Wirtschaftswissenschaften, Zeitschrift für betriebswirtschaftliche Forschung 29 (1977) 12, S. 777 ff.

Mertens, P., Computergestützte Unternehmensplanung, Informatik Spektrum 2 (1979), S. 86 ff.

Mertens, P. u.a., Verknüpfung von Daten- und Methodenbanken, dargestellt am Beispiel der Analyse von Marktforschungsdaten, in: Plötzeneder, H. D. (Hrsg.), Computergestützte Unternehmensplanung - Computer Assisted Corporate Planning, Stuttgart 1977, S. 291 ff.

Mertens, P. und Bodendorf, F., Interaktiv nutzbare Methodenbanken - Entwicklungskriterien und Stand der Verwirklichung, Angewandte Informatik (in Vorbereitung).

Möllhoff, L., Unvollkommenes Informationsnachfrage-Verhalten im Mensch-Maschine-Dialog, Stuttgart 1978.

Mojena, R., Hierarchical Grouping Methods and Stopping Rules: An Evaluation, Computer Journal 20 (1977), S. 359 ff.

Molzberger, P. und Peschanel, F. D., Entwicklung schneller Verfahren zur hierarchischen Cluster-Analyse bei großen Datenmengen, in: Kohlas, I., Seifert, O., Stähly, P. und Zimmermann, H. J. (Hrsg.), Proceedings in Operations Research 5, Würzburg 1975, S. 187 ff.

Morrison, D. G., Measurement Problems in Cluster Analysis, Management Science 13 (1967), S. B-775 ff.

Müller-Merbach, H., Tendenzen der Verwendung quantitativer Ansätze in der betriebswirtschaftlichen Forschung und Praxis, in: Müller-Merbach, H. (Hrsg.), Quantitative Ansätze in der Betriebswirtschaftslehre, München 1978, S. 11 ff.

Müller-Merbach, H. (Hrsg.), Quantitative Ansätze in der Betriebswirtschaftslehre, München 1978.

Nie, N. H., Bent, D. H. und Hull, O. H., Statistical Package for the Social Sciences, New York 1975.

Niemann, H., Methoden der Mustererkennung, Frankfurt 1974.

Niemann, H., Mustererkennung - Einführung und Überblick, Informatik Spektrum 2 (1979), S. 12 ff.

A. C. Nielsen Company (Hrsg.), Nielsen Modell SABINE, interne Firmenschrift, Frankfurt o. J.

Nollau, H. G., Ein Verfahren zur automatischen Klassifikation, in: Gesellschaft für Klassifikation (Hrsg.), Klassifikation und Erkenntnis III, Studien zur Klassifikation, Band 6, Frankfurt 1979, S. 95 ff.

Opitz, O., Neuere Entwicklungen zur Klassifikation mit qualitativen Daten, in: Gesellschaft für Klassifikation (Hrsg.), Klassifikation und Erkenntnis III, Studien zur Klassifikation, Band 6, Frankfurt 1979, S. 17 ff.

Pakin, S., APL-Handbuch, Stuttgart 1975.

Pansch, P., Erstellung und Test von Programmen zur Integration der Methodenbanksysteme MADAS und ANCA, Studienarbeit, Erlangen 1979.

Peschanel, F. D., Zu den Konsequenzen der Vernachlässigung von Korrelationen zwischen den Merkmalen bei der Anwendung von Clusteranalysen, in: Dathe, H. N., Mertens, P., Peschanel, F. D., Späth, H. und Zimmermann, H. J. (Hrsg.), Proceedings in Operations Research 6 (1976), S. 224 ff.

Pessemier, E. A., Data Quality in Marketing Information Systems, in: Farley, J. U. und Howard, J. A. (Hrsg.), Control of "Error" in Market Research Data, Lexington-Toronto-London 1975, S. 109 ff.

Pfohl, H.-C. und Rürup, B. (Hrsg.), Anwendungsprobleme moderner Planungs- und Entscheidungstechniken, Königstein 1978.

Plötzeneder, H. D. (Hrsg.), Computergestützte Unternehmensplanung - Computer Assisted Corporate Planning, Stuttgart 1977.

Rand, W. M., Objective Criteria for the Evaluation of Clustering Methods, Journal of the American Statistical Association 66 (1971), S. 846 ff.

Rickert, R., Methodenbanken - Softwaretechnologie für individuelle Planungs-, Informations- und Auswertungssysteme, Siemens-Zeitschrift 51 (1977), S. 242 ff.

Samelson, K., Entwicklungslinien der Informatik, in: Schindler, S. und Giloi, W. K. (Hrsg.), Proceedings der Jahrestagung 1978 der Gesellschaft für Informatik, Berlin-Heidelberg-New York 1978, S. 142.

Schäffer, J. B., Beschreibung und Benutzeranleitung des Wishartschen Clusteranalyse-Pakets, Bericht MD 97 der Gesellschaft für Strahlen- und Umweltforschung, München 1974.

Schaller, Karl, Eine Methodenbank zur Clusteranalyse (mit Diskussionsprotokoll), in: Gesellschaft für Klassifikation (Hrsg.), Klassifikation und Erkenntnis III, Studien zur Klassifikation, Band 6, Frankfurt 1979, S. 79 ff.

Schaller, Klaus, Verfahren zur Darstellung und Analyse von Ergebnissen der Clusteranalyse - Entwurf, Implementation und Test an Daten der Marketingpraxis, Diplomarbeit, Nürnberg 1978.

Schindler, S. und Giloi, W. K. (Hrsg.), Proceedings der Jahrestagung 1978 der Gesellschaft für Informatik, Berlin-Heidelberg-New York 1978.

Schips, B., Ein Beitrag zum Thema "Methodenbanken", Angewandte Informatik 19 (1977), S. 465 ff.

Schläger, W., Die Klassifikation von Verweildauerhäufigkeiten stationärer Patienten der Medizinischen Klinik der Universität Erlangen-Nürnberg mittels Cluster-Analyse, in: Späth, H. (Hrsg.), Fallstudien zur Clusteranalyse, München 1977, S. 81 ff.

Schmidt, R., Zur Verbindung von Modellen, Methoden und Daten bei Unternehmensplanung mit EDV, in: Müller-Merbach, H. (Hrsg.), Quantitative Ansätze in der Betriebswirtschaftslehre, München 1978, S. 95 ff.

Schneider, W., Taxonomie der Gedächtnisleistungen schwacher und normaler Rechtschreiber, in: Späth, H. (Hrsg.), Fallstudien zur Clusteranalyse, München 1977, S. 179 ff.

Schober, F., Die Rolle der Datenverarbeitung in der Ökonometrie, in: Schober, F. und Plötzeneder, H. D. (Hrsg.), Ökonometrische Modelle und Systeme, München-Wien 1978, S. 41 ff.

Schober, F. und Plötzeneder, H. D. (Hrsg.), Ökonometrische Modelle und Systeme, München-Wien 1978.

Schobert, R., Die Dynamisierung komplexer Marktmodelle mit Hilfe von Verfahren der mehrdimensionalen Skalierung, Dissertation, Nürnberg 1978.

Scholz, P., Konzeption und Realisierung eines Datenbank-dialogsystems zur mathematisch-statistischen Auswertung von Marktforschungsdaten, Diplomarbeit, Erlangen 1975.

Schuberth, V., Verknüpfung von Faktorenanalyse, Clusteranalyse und multipler Diskriminanzanalyse im Rahmen der interaktiven Methodenbank MADAS, Diplomarbeit, Nürnberg 1977.

Sheth, J. N., The Multivariate Revolution in Marketing Research, Journal of Marketing 35 (1971), S. 14 ff.

Smith, B. T. e. a., Matrix Eigensystem Routines - EISPACK Guide, Lecture Notes in Computer Science 6, Berlin-Heidelberg-New York 1974.

Späth, H., Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion, München 1975.

Späth, H. (Hrsg.), Fallstudien zur Clusteranalyse, München 1977.

Späth, H., Partitionierende Cluster-Analyse für große Objektmengen mit binären Merkmalen am Beispiel von Firmen und deren Berufsgruppenbedarf, in: Späth, H. (Hrsg.), Fallstudien zur Clusteranalyse, München 1977, S. 63 ff.

Späth, H., Numerische Erfahrungen zu heuristischen Lösungsverfahren beim Varianzkriterium in der Cluster-Analyse, Angewandte Informatik 19 (1977), S. 67 ff.

Späth, H., Computational Experiences with the Exchange Method, European Journal of Operational Research 1 (1977), S. 23 ff.

Städtler, K., Zur Methodik des Infratest-Typologie-Programms SYAT, interne Firmenschrift, München 1975.

Steinhausen, D. und Langer, K., Clusteranalyse: Einführung in Methoden und Verfahren der automatischen Klassifikation, Berlin-New York 1977.

Stumpf, H., Test und Vergleich der Programmsysteme ANCA und CLUSTAN zur Clusteranalyse, Studienarbeit, Nürnberg 1978.

Szyperski, N., Realisierung von Informationssystemen in deutschen Unternehmungen, in: Müller-Merbach, H. (Hrsg.), Quantitative Ansätze in der Betriebswirtschaftslehre, München 1978, S. 67 ff.

Topritzhofer, E., Marketing - Neue Ergebnisse aus Forschung und Praxis, Wiesbaden 1978.

Tryon, R. C. und Bailey, D. E., Cluster Analysis, New York 1970.

Vogel, F., Probleme und Verfahren der numerischen Klassifikation, Göttingen 1975.

Vogel, F., Subjektivitäten bei der Klassifikation von Einheiten, in: Brockhoff, K., Dinkelbach, W., Kall, P., Pressmar, D. B. und Spicher, K. (Hrsg.), Proceedings in Operations Research 7, Würzburg 1978, S. 105 ff.

Wendt, F., Typologische Analyse Systeme in der deutschen Marktforschung, Hamburg 1975.

Williams, W. T., Lance, G. N., Dale, M. B. und Clifford, H. T., Controversy concerning the Criteria for taxonomic Strategies, Computer Journal 14 (1971), S. 162 ff.

Winand, U., Zur Gestaltung von Modell-Methoden-Verbund-Programmen im Rahmen von quantitativen modellgestützten Problemlösungen, Angewandte Informatik 17 (1975), S. 503 ff.

Wishart, D., CLUSTAN 1C Release 1 User Manual, London 1975.

Wishart, D., Treatment of Missing Values in Cluster Analysis, in: Corsten, L. C. A. und Hermans, J. (Hrsg.), COMPSTAT 1978: Proceedings in Computational Statistics, 3. Symposium held in Leiden, Wien 1978, S. 281 ff.

Wißkirchen, P., Stand und Entwicklungstendenzen der Grafischen Datenverarbeitung, Interner Bericht der GMD, St. Augustin 1978.

Wolff, M.-R., Portfolio Management und Wertpapieranalyse vermittelt IDAMS, Heidelberg 1979.

Yao, S. B. (Hrsg.), Proceedings der Fourth International Conference on Very Large Data Bases, Berlin 1978.

Lebenslauf

Geboren	15. 12. 1949 in Naila/Oberfranken
Eltern	Karl Schaller und Anna Schaller geb. Baumann
Familienstand	verheiratet
1956 - 1960	Volksschule Naila
1960 - 1969	Schiller-Gymnasium Hof/Saale
1969 - 1974	Studium der Betriebswirtschaftslehre an der Universität Erlangen-Nürnberg Abschluß: Diplom-Kaufmann
1974 - 1975	wissenschaftlicher Angestellter der HIS GmbH, Hannover
1975 - 1979	Verwalter der Dienstgeschäfte eines wissenschaftlichen Assistenten am Lehrstuhl für Betriebswirtschaftslehre, insbesondere Betriebs- und Wirtschaftsinformatik Prof. Dr. Peter Mertens

